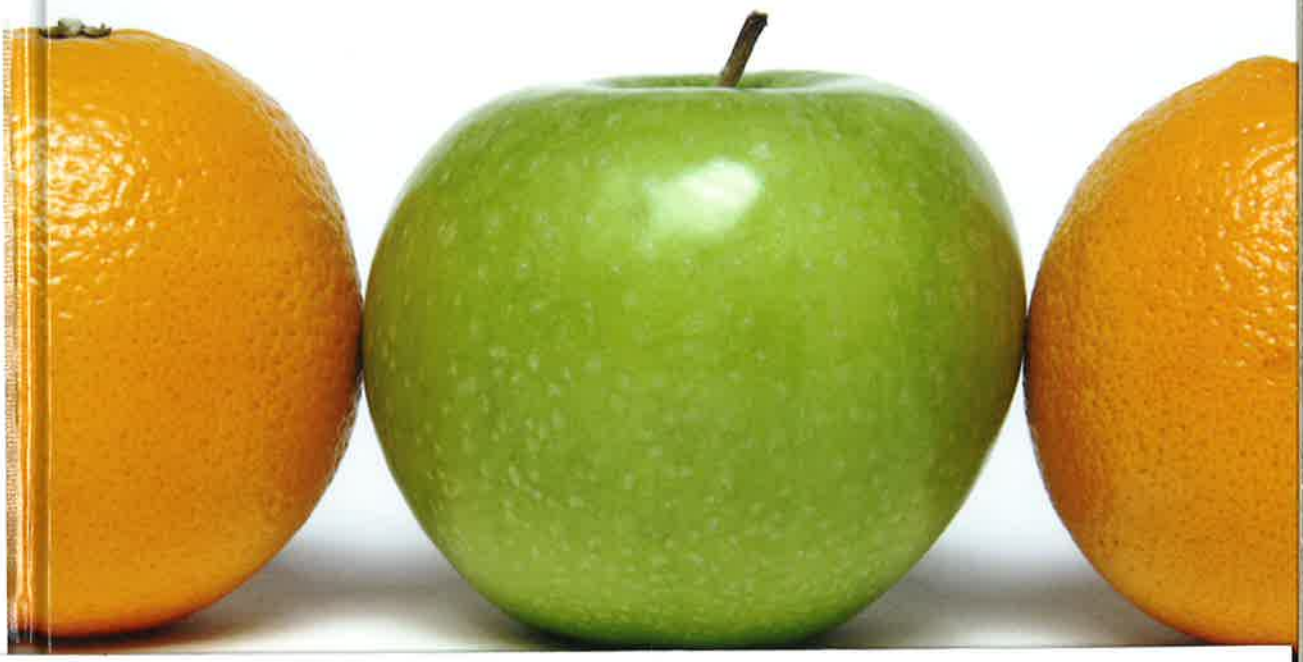
RISK ADJUSTMENT

4th EDITION FOR MEASURING

HEALTH CARE OUTCOMES

LISA I. IEZZONI, Editor



RISK ADJUSTMENT FOR MEASURING HEALTH CARE OUTCOMES

4th EDITION

LISA I. IEZZONI, Editor

HAP

AUPHA

Health Administration Press, Chicago, Illinois

Association of University Programs in Health Administration,
Arlington, Virginia

Your board, staff, or clients may also benefit from this book's insight. For more information on quantity discounts, contact the Health Administration Press Marketing Manager at (312) 424–9470.

17 16 15 14 13 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data
Risk adjustment for measuring health care outcomes / Lisa I. Iezzoni, editor. -- 4th ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-1-56793-437-3 (alk. paper)
I. Iezzoni, Lisa I.
[DNLM: 1. Outcome Assessment (Health Care)--methods--United States. 2. Risk Adjustment--methods--United States. 3. Decision Support Techniques--United States. 4. Delivery of Health Care--organization & administration--United States. 5. Models, Statistical--United States. 6. Risk Factors--United States. W 84.4 AA1]
362.1--dc23
2012017084

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984. ♾™

Acquisitions editor: Janet Davis; Project manager: Jennifer Seibert; Cover designer: Scott Miller; Layout: Fine Print, Ltd.

Found an error or a typo? We want to know! Please e-mail it to hap1@ache.org, and put "Book Error" in the subject line.

Health Administration Press
A division of the Foundation of the American College of Healthcare Executives
One North Franklin Street, Suite 1700
Chicago, IL 60606–3529
(312) 424–2800

Association of University Programs in Health Administration
2000 North 14th Street
Suite 780
Arlington, VA 22201
(703) 894–0940

BRIEF CONTENTS

DETAILED CONTENTS

DETAILED CONTENTS

PREFACE

This book introduces important conceptual and methodological considerations for designing, using, and evaluating risk adjustment methods. As I write this preface, we are at an especially critical moment for those interested in risk adjustment. Policymakers are planning and implementing initiatives employing risk adjustment to control health care costs, expand access to clinically complex populations, and improve the safety and quality of care. Data sources supporting risk adjustment are undergoing radical transformations, including the growing deployment of electronic health records and new technological approaches to obtain information directly from patients, both of which should expand the applicability of risk adjustment to a wide range of purposes and enrich its validity. Thus, the pressures on—but also the promise of—risk adjustment have never been greater.

However, this moment is also one of tremendous uncertainty. At the end of March 2012, the US Supreme Court heard oral arguments in the lawsuit challenging the Patient Protection and Affordable Care Act (ACA), enacted just over two years ago. Many of the initiatives involving risk adjustment cited earlier relate directly to various ACA provisions, and suspension of ACA mandates would likely affect the timing and nature of certain initiatives significantly. The Court's oral arguments escalated speculation about the ACA's fate to a fever pitch among health policy observers; expectations are that a single vote will determine the decision. Nonetheless, if the Court overturns the ACA or curtails its reach, pressing needs to control US health care costs, treat complex patients, and ensure quality will not disappear. The demand for intelligent risk adjustment will continue to grow, even if the ACA's effects ebb.

As have prior editions, this fourth edition focuses on basic principles and methodologies relating to risk adjustment. We do not attempt to provide up-to-date information on current methods: They are changing too rapidly, and new approaches are undergoing development and testing, making it impossible for a textbook to capture the cutting edge of the field. Nonetheless, we do venture into some areas undergoing rapid transformation. We update our data discussions, adding early insights relating to health information technologies. We discuss newer statistical methods that the field is increasingly using but that require further testing to be fully understood.

We briefly address such topics as conducting surveys, steps for developing risk adjustment methods, measurement of validity and reliability, the use of risk adjustment in managing care, and issues relating to special populations, including children, persons with mental health conditions, persons with disabilities, and persons requiring long-term support services. We anticipate this book will have a multidisciplinary audience, some readers concentrating only on certain chapters. Because chapters may be used selectively, we repeat key concepts throughout the book. The text emphasizes issues most relevant to risk adjustment and does not duplicate the detailed technical discussions found in statistical and methodological textbooks.

Given the proprietary nature of many risk adjustment methods, we must declare potential conflicts of interest. Most of us have received research funding to investigate risk adjustment at one or more points during our careers. I served for many years as a clinical co-investigator with Arlene S. Ash, as she and other colleagues developed and refined the Diagnostic Cost Group (DCG) methodology under cooperative agreements and contracts with the Medicare agency. Arlene cofounded DXCG, the Boston-based company that supported and marketed the DCGs, but she gave up her equity interest when the firm was sold to Verisk Health (Waltham, Massachusetts). Arlene now serves as senior scientist at Verisk Health, which continues to develop and market DCG-based products. I have not participated in DCG commercial activities. Under a grant from the Agency for Healthcare Research and Quality, Jennifer Daley and I, along with others, developed the Complications Screening Program (CSP), which is mentioned at several points in this book. Neither she nor I have personal financial or intellectual interests in commercial or other applications of the CSP.

For instructors who use our book in their courses, we have developed discussion questions and supplemental material. Thanks go to all the authors, but in particular to Arlene Ash and Michael Shwartz, for their additional work to create these useful instructor's resources. These educational materials, as well as PowerPoint slides of all the exhibits in the book, are available by sending an e-mail to hap1@ache.org.

We extend special thanks to contributors to previous editions, who graciously allowed us to borrow from their contributions, notably John S. Hughes, MD, and Richard C. Hermann, MD. The previous authors of chapters 8 and 9 (Jennifer Daley, Arlene Ash, and Michael Shwartz) relinquished their roles to me; I am in their debt. Amy J. Wint, my administrative assistant and research coordinator, diligently and efficiently dealt with hundreds of bibliographical citations and assisted with numerous other production tasks.

Finally, most of us wrote our contributions to this book largely during our free time. For some spouses and partners, this was the fourth time they had to endure weekends and evenings, stretching over weeks and sometimes many months, of intensive preoccupation with THE BOOK. We reward

them for this forbearance with a promise that I now put into writing (the previous promise was made only verbally and thus more easily broken): We promise never to do this again! This fourth edition will be our last. We are deeply grateful to readers who, since 1994, have thanked us for this book and said many kind words about it. Through it we share what we have learned during nearly 30 years of working in risk adjustment, trusting the next generation of researchers to move the field leaps and bounds beyond where we leave it today.

Lisa I. Iezzoni
April 18, 2012

CHAPTER

1 REASONS FOR RISK ADJUSTMENT

Lisa I. Iezzoni

In the early twenty-first century United States, health care is poised, somewhat uneasily, in a paradoxical place. On the one hand, we face the promise of stunning scientific breakthroughs from unraveling the human genetic code and greater insights into the molecular basis of disease. Someday, we might truly personalize medicine by prescribing therapies on the basis of genetic indications that a patient's disease will respond, or we might manipulate interventions at the molecular level with nanotechnology. Extraordinary technical advances have occurred. With laparoscopic surgery, surgeons can remove tumors through tiny incisions. Sophisticated imaging technologies are deepening our understanding of the human brain. Through studies of primitive species like fruit flies, we are learning about the process of aging and ways to arrest what previously seemed an inevitable degenerative progression. Some scientists are talking about the possibility of extending human life.

On the other hand, we confront the possibility that younger generations of Americans will have worse health than their parents do—a phenomenon that has never occurred in the nation's history. High rates of obesity and overweight drive much of this fear, but so too do other factors, including individual choice (e.g., physical inactivity) and powerful societal and environmental forces, such as deep and persistent poverty, especially among children. Other critical factors are individuals' inability to pay for health care and lack of health insurance. Uninsurance is strongly linked to worse health (McWilliams 2009). In 2010, the number of Americans without health insurance increased to 256.2 million, up from 255.3 million in 2009, although the percentage of uninsured individuals remained statistically constant at approximately 16 percent across the two years (DeNavas-Walt, Proctor, and Smith 2011, 23). Health insurance coverage and thus financial access to care can substantially improve health, even among socioeconomically disadvantaged populations (Finkelstein et al. 2011).

Political maelstroms and crosscurrents stymied efforts to extend health insurance coverage to all Americans for nearly a century. In the 1930s, Franklin Delano Roosevelt and his associates envisioned "a universal right to health care secured by a national compulsory insurance plan"—a legacy left to Harry Truman, who declared the failure to achieve this vision "the great

lost cause of his life" (Blumenthal and Morone 2009, 15). President after president tried and failed to pass even modest reforms to expand health insurance. The critical—and enormous—exception was the 1965 enactment of Medicare (the federal health plan for elderly and disabled individuals[1]) and Medicaid (the joint federal-state program for low-income persons in specified eligibility categories).[2] Subsequent efforts to ensure that all Americans have health insurance continued to fail, sometimes spectacularly, for more than four decades.

In 2009, an unlikely alignment of political stars permitted passage of the Patient Protection and Affordable Care Act (ACA, P.L. 111-148 with its companion Health Care Education and Reconciliation Act of 2010, P.L. 111-152), signed on March 23, 2010, by President Barack Obama. As perhaps its main goal, ACA aims to ensure that the majority of Americans have health insurance coverage. According to Congressional Budget Office (CBO 2011) projections, under ACA provisions roughly 95 percent of legal, nonelderly US residents will have health insurance in 2021, compared with a projected 82 percent without the law.

Within a year of the law's enactment, however, the political stars realigned dramatically, spurred partially by antipathy toward the ACA. On January 19, 2011, the US House of Representatives passed H.R. 2, the Repealing the Job-Killing Health Care Law Act. Lawsuits questioning the constitutionality of key ACA provisions wound through lower federal courts before landing in the US Supreme Court. The Supreme Court should issue its decision about the constitutionality of the individual mandate (requirements that individuals have health insurance) and several other ACA provisions by June 2012. Furthermore, as Blumenthal (2011, 2047) observed, "The 2012 election will be the most important in the history of our health care system because it will determine whether the [ACA] is implemented or repealed." Thus, at the time of this writing, the ACA's future remains politically and legally uncertain, exacerbating the uneasy equipoise of the US health care system.

Beyond insuring the uninsured, unsustainable cost growth and questionable quality confound US health care. In addition to its goals of insuring almost all Americans, ACA has numerous other aspirations, including "bending the cost curve" (i.e., reducing ever-escalating costs of care) and improving the quality of care (e.g., by enhancing care coordination and linking health care expenditures to the value of care). National health care expenditures in 2012 are projected to top $2.85 trillion (17.2 percent of the gross domestic product [GDP]) and to rise to $4.6 trillion (19.6 percent of GDP) by 2019 (Blumenthal 2011; CMS 2010a). In March 2011, the CBO estimated that various ACA provisions would decrease Medicare and Medicaid spending sufficiently to reduce federal deficits by $210 billion over the 2012–2021 period (CBO 2011).

Compounding problems with cost and access are concerns about what the nation is purchasing for its trillions of health care dollars. In its seminal report *Crossing the Quality Chasm,* the Institute of Medicine (IOM 2001a, 2) concluded, "As medical science and technology have advanced at a rapid pace . . . the health care delivery system has floundered in its ability to provide consistently high-quality care to all Americans." A decade later, rates of errors and safety problems in health care remain unacceptably high (AHRQ 2011b). Striking variations across small geographic areas, first noted nearly four decades ago (Wennberg and Gittelsohn 1973), also persist; expenditures per Medicare beneficiary vary widely across regions. In 2008, mean Medicare expenditures per individual beneficiary ranged from less than $7,000 in places such as Rapid City, South Dakota ($6,264), Honolulu ($6,653), and Portland, Oregon ($6,971), to more than $12,000 per beneficiary in such locations as Miami, Florida ($15,571), McAllen, Texas ($14,529), and Monroe, Louisiana ($12,027) (Skinner, Gottlieb, and Carmichael 2011, 3). Fragmented, poorly coordinated, and ill-designed systems of care are particularly inadequate for the needs of people with chronic conditions, who consume the most resources.

From a population perspective, large disparities exist in health and health care between privileged and disadvantaged subgroups, such as racial and ethnic minorities and persons with disabilities (AHRQ 2011a). The extent to which the health care system contributes to disparities in health among individuals and populations is unclear, but we do know that social and environmental factors beyond what is traditionally considered health care are critical determinants of health. The *Healthy People 2020* initiative, which sets the road map for improving the health of Americans in the decade ahead, addresses determinants of health head-on (Secretary's Advisory Committee 2008, 9):

> Responsibilities for promoting healthful environments go beyond the traditional health care and public health sectors. Changes in social environments, physical environments, and policies can affect entire populations over extended periods of time and help people to respond to individual-level interventions. Policies that can increase the income of low income persons and communities (e.g., through education, job opportunities, and improvement in public infrastructure) may improve population health. Reducing inequalities in the physical environment (e.g., access to healthful foods, parks, and transportation) can also improve key health behaviors and other determinants, thereby helping to meet numerous health objectives.

Given this complex and multilayered context, numerous compelling arguments support the need to examine and better understand the health care delivery system. For individual patients, more information could assist decision making not only about the benefits of specific clinical interventions but also about which providers offer high-quality care. Purchasers,

payers, and policymakers need to identify providers and health care systems meriting remediation or improvement. Devising ways to measure health care value—an amalgam of quality and costs—can assist efforts to enhance efficiency. As ACA reforms begin, continued measurement will be essential to track successes and failures, practices to emulate, and practices requiring heightened vigilance. As suggested by these varied purposes, productive examination of the health care system requires comparisons across different units of observation: subgroups of patients, treatments, providers, delivery systems, health plans, and populations. The goal is to answer questions such as the following:

- Which treatments are most effective?
- Which providers and approaches produce the best outcomes?
- Which health plans, delivery systems, and providers are most efficient?
- What is the best way to set payments for accountable care organizations and other delivery systems?
- Which models of health care delivery provide the highest-value care to patients and purchasers?
- How should care, providers, delivery systems, and health plans be monitored to ensure high quality and efficiency?

Purpose of Risk Adjustment

Before meaningful comparisons of outcomes can be made across different patients, treatments, providers, delivery systems, health plans, and populations, we must account for patient-associated factors. Risk adjustment would be unnecessary if individuals were randomly assigned to comparison groups, but they are not. Many factors affect how persons find care, ranging from specific health concerns to financial considerations to geographic location to preferences and expectations for health outcomes and services.

The mix of persons treated by different clinical interventions, providers, or health plans varies. These differences have consequences. Higher-risk persons typically generate greater costs, even with efficient providers or health plans, than do lower-risk persons. Persons with complex illnesses, multiple coexisting diseases, or other significant risk factors generally develop more complications and experience worse outcomes, even with excellent care, than do healthier individuals. Risk adjustment aims to account for differences in intrinsic health risks that patients or populations bring to their health care encounters. When determining costs or comparing outcomes, it "levels the playing field," ensuring that "apples are compared to apples, not to oranges"—like-to-like comparisons.

Risk adjustment is essential to examining outcomes in the real world. In the artificial environment of randomized controlled trials (RCTs), experts administer treatments according to tightly specified protocols in closely monitored settings. Randomization presumably sorts patients into comparison groups regardless of their baseline risks. It theoretically produces treatment (case) and nontreatment (control) groups with similar risk factors. RCTs thus yield the "gold-standard" evidence of treatment efficacy. In examining the routine care practiced daily throughout communities, however, the focus shifts to quantifying effectiveness—the outcomes of care as it is delivered in the real world. Risk adjustment is integral to calculating the "algebra of effectiveness," which recognizes that health care outcomes are a complex function of intrinsic patient attributes and other factors:

$$\text{Outcomes} = f(\text{intrinsic patient-related risk factors, treatment effectiveness, quality of care, random chance}).$$

The specifics of this equation vary depending on the outcome of interest and the range of patient attributes that must be considered. These equations can become complex and far reaching; many attributes, clinical and nonclinical, contribute to health-related risks. More than three decades of intensive research have produced credible risk adjustment methods for outcomes in widely divergent contexts, such as in

- predicting in-hospital mortality for children and adults treated in intensive care units (ICUs);
- predicting in-hospital mortality and postoperative complications for coronary artery bypass grafting and other major surgeries;
- examining patient-reported experiences with health care or health plans;
- setting prospective payment levels for specific episodes of care (Exhibit 1.1), such as acute care hospital reimbursement based on Diagnosis-Related Groups (DRGs) and nursing home payments using Resource Utilization Groups (RUGs); and
- setting capitation payment levels for managed care organizations (MCOs) enrolling Medicare beneficiaries, Medicaid recipients, and other members.

Existing risk adjustment methods are widely used by private and public payers, governmental and voluntary oversight bodies (e.g., for provider accreditation), quality improvement organizations, health data organizations, media (e.g., magazines producing their "best hospitals" issues), clinicians, health services researchers, and others. Risk adjustment methods remain underdeveloped for many critical aspects of care, such as routine outpatient settings, nonsurgical inpatient care, and specific subpopulations (e.g., children

EXHIBIT 1.1 Risk Adjustment Methods Used by Medicare to Set Payments

Classification System	Covered Setting	Year[a]
Diagnosis-Related Groups (DRGs)	Acute care hospital payment per inpatient discharge	1984
Medicare Severity DRGs (MS-DRGs)[b]	Acute care hospital payment per inpatient discharge	2008
Medicare Severity Long-Term Care DRGs (MS-LTC-DRGs)	Long-term care hospital payment per inpatient discharge	2006
Ambulatory Payment Classification (APC)	Hospital payment for outpatients discharged from hospital emergency rooms or clinics or transferred to other facilities	2000
Resource Utilization Groups (RUGs)	Skilled nursing facilities	1998
Home Health Resource Groups (HHRGs)	Home health agencies	2001
Case-Mix Groups (CMGs)	Inpatient rehabilitation facilities	2002
CMS Hierarchical Condition Categories (CMS-HCCs)	Capitated payments for Medicare managed care organizations	2004
Drug Hierarchical Condition Categories (RxHCCs)	Medicare Part D payments for medications	2006

[a]Year in which payment method was instituted by Medicare. Most years represent fiscal years (i.e., year starts October 1 of prior calendar year).
[b]Medicare severity-adjusted DRGs replaced standard DRGs for setting prospective hospital payments in fiscal year 2008.

not in ICUs, persons with mental health conditions or disabilities). Although risk adjusters for setting capitation payments are substantially better than standard demographic approaches, questions remain about whether supplementary policies (e.g., partial capitation) are needed to improve fairness of payment for particularly costly enrollees.

Thus, the devil of risk adjustment is in the details. As described in Chapter 2, one cannot begin risk adjustment before defining basic terms: risk of what? Failing to account adequately for patients' risks can prove embarrassing. A prime example was the first public release in March 1986 of Medicare's hospital mortality figures by the Health Care Financing Administration (HCFA, renamed the Centers for Medicare & Medicaid Services [CMS] in 2001) (Brinkley 1986). According to HCFA, 142 hospitals had significantly higher death rates than predicted, whereas 127 had significantly lower rates. At the institution with the most aberrant death rate, 87.6 percent of Medicare patients died compared to a predicted 22.5 percent. This facility, however, was a hospice caring for terminally ill patients. HCFA's risk adjustment model had not adequately captured patients' risks of death.[3]

Goals of This Book

As do prior editions, this book examines conceptual and methodological issues raised by risk adjustment when comparing outcomes of care, including costs, clinical outcomes, and patient-centered outcomes, broadly defined, in various health care settings. As noted in Chapter 2, we do not review existing risk adjustment methods; the field is too broad and constantly evolving. Instead we concentrate on basic methods and principles that should apply generally to risk adjustment, regardless of setting or purpose. We briefly address such topics as conducting surveys, measuring validity and reliability, general linear and logistic regression modeling, propensity scores and instrumental variables, and hierarchical modeling. Discussion of each topic emphasizes issues most relevant to risk adjustment and does not duplicate detailed technical discussions found in statistical and methodological textbooks. We use examples from seminal studies, even if they occurred years ago, when the studies demonstrate core concepts.

Risk adjustment will be center stage as health care reform unfolds. Title I, Part V of the ACA, "Reinsurance and Risk Adjustment," describes requirements by January 2014 to develop state-specific methods to derive payment levels for group health plans caring for high-risk individuals and for setting payments that account for enrollees' health risks. Part V, Section 1343 addresses risk adjustment specifically and mandates differential payments to plans determined to have low versus high actuarial risks. Title III, "Improving the Quality and Efficiency of Health Care," contains many references to risk adjustment for quality measurement and reporting. For instance, Title III, Part 1, "Linking Payments to Quality Outcomes Under the Medicare Program," raises concerns about risk adjustment in producing the quality metrics that Medicare will employ for its "Hospital Value-Based Purchasing Program" (Sec. 3001).

We do not try to anticipate where risk adjustment developments prompted by the ACA will ultimately lead. By the time readers open this book, methodologically the field may have changed considerably from where we are now, or perhaps recurrent challenges to risk adjustment, described in ensuing chapters, will have stymied efforts to achieve ACA goals (e.g., value-based purchasing initiatives). Even if the ACA is repealed or does not move forward as planned, risk adjustment will remain an important consideration in future efforts to control costs while ensuring and monitoring health care quality.

Historical Precedents

Historical precedents for examining outcomes of health care are noteworthy not only because of the effort taken to compile and compare outcomes data but, more important, because they involved vigorous endeavors to determine what

caused variations and to use this knowledge to improve care. Risk adjustment figures prominently in these historical examples. Many motivations and consequent controversies uncannily anticipated today's discourse (Iezzoni 1996).

For centuries, Britain gathered data on population death rates, primarily to track epidemic illness. Overwhelmed by deaths caused by plagues, royal authorities initiated weekly "bills of mortality" in the early 1500s (Walker 1929). In the late eighteenth century, profound social upheavals of the industrial revolution spurred this interest in tracking mortality statistics. As populations moved from the countryside to increasingly congested industrial centers, statistics depicted worsening public health (Iezzoni 1996). By the 1830s, civic and business leaders had founded statistical societies throughout England, aiming to quantify the effects of social changes. The archetypal member was "a liberal Whig, Unitarian, reform-minded" (Eyler 1979, 14). These early Victorian statisticians viewed facts as the scientific impetus for political change.

English hospitals, which were primarily charitable institutions serving the poor, had independently accumulated patient statistics since the 1600s. In the mid-nineteenth century, even England's registrar-general had trouble determining which facilities (or parts of facilities) were hospitals as opposed to workhouses; in 1861, workhouses contained 81 percent of beds for physically ill persons (Pinker 1966). Hospitals compiled death statistics to show wealthy benefactors that their charity contributions were being used wisely and to encourage new subscribers and donations. As today, individual philanthropists and organizations funding hospitals wanted proof they were getting their money's worth. As noted in an 1863 report to the medical officer of the Privy Council, "The public as a rule still look to the death-rates of hospitals as the best indication of their relative healthiness" (Bristowe and Holmes 1864, 512).

Improving and Monitoring Outcomes

In 1863, Florence Nightingale (1820–1910; Exhibit 1.2) published the third edition of her *Notes on Hospitals*, recommending fundamental changes in the configuration, location, and operation of hospitals to reduce deaths caused by unsanitary conditions. Seven years earlier, Nightingale had returned from Crimean War service at British military hospitals as perhaps the first wartime celebrity ever created by the news media (Cohen 1984). Crafted by a *Times* correspondent, her image as the lone lady nursing sick soldiers lit by her handheld lamp earned Nightingale an admiring lifelong audience. This gentle, ministering angel persona, however, belied her tough-minded, laser-focused administrative acumen: In 1855, six months after arriving at Barrack Hospital in Scutari, Turkey, she cut military hospital death rates from 42.7 percent to 2.2 percent (Cohen 1984).

After returning home, Nightingale continued targeting military installations, but needing statistical help, she turned to William Farr (1807–1883),

EXHIBIT 1.2 Florence Nightingale and Ernest Amory Codman

Source: The Boston Medical Library in the Francis A. Countway Library of Medicine. Used with permission.

a physician and prominent social reformer who had conducted analyses for the registrar-general since 1838. In 1856, they made a pact: Farr would assist her with army reforms, and Nightingale would aid his efforts to reduce civilian deaths (Eyler 1979). In her 1863 *Notes on Hospitals*, Nightingale concentrated primarily on civilian hospitals.

Farr and Nightingale viewed high hospital death rates as incontrovertible proof of the dangers of mid-nineteenth-century urban hospitals; Exhibit 1.3, taken from her book, shows the deaths at "106 principal hospitals of England" in 1861. Most startling was the 90.84 "mortality per cent. on inmates" at 24 London hospitals, taken verbatim from Farr's *24th Annual Report of the Registrar-General.* These figures led Nightingale (1863, 4) to question the value of these inner-city hospitals:

> Facts such as these (and it is not the first time that they have been placed before the public) have sometimes raised grave doubts as to the advantages to be derived from hospitals at all, and have led many a one to think that in all probability a poor sufferer would have a much better chance of recovery if treated at home.

Nightingale warned that such figures did not adequately capture patients' risks, observing that, at a minimum, one must consider differences across hospitals in patients' ages and "state of the cases on admission" (Nightingale 1863, 2). Despite these caveats, she observed that death rates were lower at facilities with better sanitation and less crowding and at sites away from raw sewage and urban congestion. These observations supported Nightingale's theory about miasmas—noxious vapors spreading disease—and led her to

EXHIBIT 1.3
Mortality Per Cent. in the Principal Hospitals of England: 1861

	Number of Special Inmates on the 8th April, 1861.	Average Number of Inmates in each Hospital.	Number of Deaths registered in the Year 1861.	Mortality per Cent. on Inmates.
In 106 Principal Hospitals of England	12709	120	7227	56·87
24 London Hospitals	4214	176	3828	90·84
12 Hospitals in Large Towns ...	1870	156	1555	83·16
25 County and Important Provincial Hospitals	2248	90	886	39·41
30 Other Hospitals	1136	38	457	40·23
13 Naval and Military Hospitals ...	3000	231	470	15·67
1 Royal Sea Bathing Infirmary (Margate)	133	133	17	12·78
1 Dane Hill Metropolitan Infirmary (Margate)	108	108	14	12·96

Source: Nightingale (1863).

propose changes in ward configuration, sanitation, and hospital location that ultimately helped reduce hospital mortality. She introduced fresh air, light, and ample space into hospitals and apportioned patients to separate pavilions.

Nightingale continued to argue that compilation and dissemination of hospital outcome statistics were critical to understanding and improving care. With an eerily modern ring, she lamented the state of this endeavor (Nightingale 1863, 2):

> Accurate hospital statistics are much more rare than is generally imagined, and at the best they only give the mortality which has taken place in the hospitals, and take no cognizance of those cases which are discharged in a hopeless condition, to die immediately afterwards, a practice which is followed to a much greater extent by some hospitals than by others. We have known incurable cases discharged from one hospital, to which the deaths ought to have been accounted, and received into another hospital, to die there in a day or two after admission, thereby lowering the mortality rate of the first at the expense of the second.

Nightingale (1863, 4) emphasized that numbers should not focus only on mortality, stating, "If the function of a hospital were to kill the sick, statistical comparisons of this nature would be admissible." She asserted that hospitals aim to restore health, so statistics should concentrate on recovery. Nevertheless, 140 years after Nightingale's observations, information on patients' health following medical encounters is rarely available.

Publication of Nightingale's *Notes on Hospitals* unleashed several months of acerbic public debate between Farr and his methodological critics.

The testy tone and issues raised parallel today's controversies surrounding releases of performance reports on physicians and hospitals. This story therefore resumes in Chapter 18.

Analyzing End Results

The most articulate early-American proponent of monitoring outcomes of care was Ernest Amory Codman (1869–1940; Exhibit 1.2), a Boston surgeon (Berwick 1989; Donabedian 1989; Mulley 1989; Neuhauser 1990). Although perhaps apocryphal, the story of how Codman became interested in monitoring outcomes offers insight into not only his character but also his future methods. Codman and his Harvard Medical School classmate Harvey Cushing (1869–1939), who became a renowned neurosurgeon, served together as clerks at the Massachusetts General Hospital (Neuhauser 1990). The medical students' role was to administer anesthesia to surgical patients. After being anesthetized, Cushing's first patient vomited and died. This event troubled Cushing, but the hospital's senior surgeon was unconcerned, stating that such deaths were fairly common. Cushing and Codman challenged each other to compare their patients' outcomes during the clerkship. Both students maintained intraoperative records on each anesthetized patient; Codman's charts graphed the patient's pulse and respirations every five minutes. The winner of this challenge is unclear. In 1920, Cushing remembered that Codman had won, but in 1939, he wrote that Codman had lost (Neuhauser 1990). These records were the first intraoperative anesthesia charts. Today, maintenance of such charts is standard practice.

This experience initiated Codman's lifelong interest in (some might say obsession with) surgical outcomes. He willingly compared his results with those of others, albeit acknowledging that "[c]omparisons are odious, but comparison is necessary in science. Until we freely make therapeutic comparisons, we cannot claim that a given hospital is efficient, for efficiency implies that the results have been looked into" (Codman 1934, xxiii). Codman's unique contribution was the "end-results idea"—an attempt to link specific interventions to their effects on patients. He stated that this idea

> . . . was merely the commonsense notion that every hospital should follow every patient it treats, long enough to determine whether or not the treatment has been successful, and then to inquire "if not, why not" with a view to preventing similar failures in the future. (Codman 1934, xii)

For operative patients, follow-up could mean monitoring for years after the surgery.

Codman tried putting his end-results idea into practice at Massachusetts General Hospital, tracking down patients a year after surgery and examining them. Some of his surgical colleagues viewed these activities as extreme.

Nevertheless, Codman (1918, 137) argued that this work was essential to improving the quality of care:

> So I am called eccentric for saying in public: that Hospitals, if they wish to be sure of improvement,
>
> - Must find out what their results are.
> - Must analyze their results, to find their strong and weak points.
> - Must compare their results with those of other hospitals. . . .
> - Must welcome publicity not only for their successes, but for their errors. . . .
>
> Such opinions will not be eccentric a few years hence.

Discouraged about the prospects for fully implementing the end-results idea at Massachusetts General Hospital, in 1911 Codman opened his own ten-bed hospital on Pinckney Street in Boston's Beacon Hill, where two dozen other surgeons, including Cushing, assisted him (Neuhauser 1990). Codman completely instituted his end-results tracking system, and he paid the Thomas Todd Co. to print annual volumes documenting the outcomes of each case the hospital treated. When he felt a treatment error or other failure occurred, he categorized the cause (Codman 1918)—for example:

- Errors due to lack of technical knowledge or skill
- Errors possibly due to lack of judgment
- Errors due to lack of care or equipment
- Errors due to incorrect diagnosis
- Cases in which the nature and extent of the disease were the main cause of failure
- Cases who refused to accept treatment

Thus, Codman linked specific outcomes to specific interventions or errors in a way that truly informed and went far beyond the information provided in today's performance reports on physicians and hospitals.

Codman fought doggedly against what he saw as laxity in the medical establishment, as suggested by the dedication of the circa 1918 publication of his hospital's end results:

> This Volume is Dedicated to
>
> RICHARD C. CABOT
>
> because I respect his motives, admire his courage and energy,
> but heartily disapprove of some of his opinions and methods,
> for he seems to want to reform the bottom of the
> profession, while I think the blame
> belongs at the top.

In 1914, Codman resigned from Massachusetts General Hospital to protest the seniority system of promotion, seeing it as antithetical to the end-results idea. The day his resignation was accepted, he reapplied, asking to be appointed surgeon-in-chief because the end results of his cases for the last decade were better than those of other surgeons. The hospital ignored his request, but by 1916 it had abandoned its seniority system of promotion. Nonetheless, Codman's candor made him unpopular. With few referrals, his Pinckney Street hospital closed in 1918.

In discussing the end-results idea, Codman (1918, 93) clearly recognized the concept of risk:

> For the man who practices surgery, there are two kinds of mortality—chance and intentional.
>
> Chance mortality is the kind which occurs unexpectedly, and which no amount of foresight can prevent. It is caused by unanticipated Calamities or Catastrophes. Death from pulmonary embolism is a good example. . . . Is it not possible to determine what this percentage of danger is, just as easily as it is to compute fire risk? . . .
>
> Intentional mortality is incurred by the chief surgeon when he attempts cases in which the condition is acknowledged to be grave. It is speculative—like gambling against known chances in a game in which skill, judgment, and luck all count.

Surgeons in Codman's era, however, had a different perspective than that of surgeons today. Nowadays, publicly releasing mortality or other outcomes information raises concerns that surgeons will avoid difficult cases, fearing that poor results will be held against them. In Codman's day (1918, 105),

> . . . a certain number of deaths are necessary to the surgeon in his business. A surgeon whose cases always get well gets no reputation for "nerve." It is said that he will never take a chance when he ought to do so. A surgeon must be "fearless" and "bold," and the only way he can prove that he is, is by a death now and then in his practice.

Codman's (1918, 106) attitude toward compensation is also rare today:

> Shall I say in the future?:
>
> 1. You are too bad a risk; go to a first-class surgeon.
> 2. You are a bad risk; I must double my usual fee.
> 3. You are a bad risk; you need not pay unless you live.
>
> All are logical. I like the last best.

Conclusion

Both Nightingale and Codman viewed outcomes information as a means to the end of improving patient outcomes and quality of care. Their work taught an important lesson: Simply knowing rates of events is insufficient;

knowing why the events occurred is essential. Risk adjustment aims to isolate one potential cause: intrinsic patient attributes that increase the likelihood of poor outcomes.

Although this book concerns measurement of risk, we acknowledge that risk adjustment is only a first step in examining the costs and quality of health care. It is simply a tool for identifying what is really important, such as inefficiencies and substandard quality, and for paying fairly for care. Other methods and approaches are required for the more important next step of developing approaches to understand and improve outcomes and the value of care.

Notes

1. In 1972 President Richard Nixon granted Medicare eligibility to persons under age 65 with disabilities. To be eligible, individuals must have received cash benefits from Social Security Disability Insurance for 24 months. A number of other categories of individuals are also eligible for Medicare, including persons with end-stage renal disease and certain types of dependents of Medicare beneficiaries.
2. President Lyndon Johnson signed the Medicare and Medicaid legislation on July 30, 1965, at the Truman Library in Independence, Missouri, with former President Harry Truman sitting at his side.
3. In June 1993, newly appointed HCFA administrator Bruce Vladeck halted production of the Medicare hospital mortality reports, concerned that they unfairly penalized inner-city public institutions, which routinely had higher-than-expected death rates. The administrative data used for the HCFA reports did not allow adjustment for critical risk factors associated with poverty and medical indigence. CMS reinstituted reports of hospital mortality rates in 2007 but only for certain conditions.

CHAPTER

2

GETTING STARTED AND DEFINING TERMS

Lisa I. Iezzoni

The first step in risk adjustment is defining terms. Many clinicians erroneously assume that other health care professionals define terms such as *risk* and *case mix* the same as they do. In the 1980s, non-clinicians in the health care delivery and health policy arenas adopted similar terms, prompted by the introduction of Medicare's inpatient prospective payment system based on Diagnosis-Related Groups (DRGs). *Risk adjustment* joined other poorly defined but oft-used words and phrases, such as *complexity*, *severity*, *intensity*, and *health status*, used not only by clinicians and researchers but also by others involved in health care delivery. Researchers, payers, policymakers, managers, regulators, quality measurement and improvement experts, performance profilers, and health insurance actuaries assign fundamentally different meanings to these terms. Different definitions of these words complicate discussions about critical health care issues.

Throughout this book, we use *risk adjustment* broadly to mean accounting for patient-related factors before examining outcomes of care, regardless of the context. To define and devise appropriate risk adjustment strategies, however, we must be specific. We start with four major questions:

1. Risk of what outcome?
2. Over what time frame?
3. For what population?
4. For what purpose?

Other questions soon follow, such as: Considering what risk factors? Using which data source? Employing which analytic methods? This chapter sketches answers to these questions, and the remaining chapters fill in more details.

We cannot exhaustively review risk adjustment methods; the field is too large and grows continuously. Furthermore, details of risk adjusters change as developers revise and update their methods. Some risk adjustment methods are widely used for highly public purposes, such as establishing hospital reimbursement levels, setting capitated payments to health plans, or producing provider performance profiles posted on websites. Many more have been developed for disease-specific research projects or quality measurement initiatives. Numerous

commercial methods, primarily intended for managerial or administrative purposes, are also available for use; their complete logic is rarely open to external scrutiny. Therefore, we use examples from seminal studies, as well as our own research, to examine risk adjustment methods that may or may not be the most up-to-date versions but nevertheless illustrate critical points that will always be relevant. As described in Chapter 8, development of risk adjusters *de novo* is complicated and often frustrating. We generally recommend taking methods "off the shelf" if their attributes match a project's goals (or the policy context) reasonably well.

Most of our examples involve risk adjusters that have been reported in the peer-reviewed literature or are described on public websites (e.g., methods the Centers for Medicare & Medicaid Services [CMS] uses to risk-adjust hospital mortality rates for the Hospital Compare performance measures). The glossary at the end of this book includes a selection of major risk adjustment methods (and their acronyms). Readers interested in using specific methods should not rely on descriptions provided here, as they may be out of date. When choosing a risk adjustment method, users should seek one with transparent logic—that is, information about the method's inner workings is completely open to outside examination. Only with complete access to the algorithm can users judge whether a particular risk adjustment method is fully appropriate for the intended purpose and understand completely the results it generates.

Risk of What?

As stated in Chapter 1, the notion of risk permeates daily life. The word portends negative consequences; the *American Heritage Dictionary* (2000) defines it as "the possibility of suffering harm or loss; danger" and "a factor, thing, element, or course involving uncertain danger; a hazard." However, the term *risk adjustment* is meaningless without identifying the outcome being risked. Answers fall broadly into three camps:

- Clinical outcomes of care, such as deaths, complications, physical functional status, and mental health
- Resources used or required, such as costs for a hospitalization, a year of care, or lengths of hospital stays
- Patient-centered outcomes, such as patients' reports that care met their preferences and expectations

Two brief vignettes demonstrate differences among definitions of risk. An adenocarcinoma of the lung was detected on a routine chest radiograph of Mr. A taken during an employment physical examination. Radiographically,

the tumor looked like an isolated lung nodule. A needle biopsy of the nodule identified lung cancer. Mr. A underwent further diagnostic testing, including a positron emission tomography scan and a magnetic resonance imaging scan of his head, to determine whether the cancer had metastasized elsewhere. Finding no metastases, Mr. A's oncologist performed a lobectomy, a major operation to remove the tumor; Mr. A experienced no complications. Mr. A, a nonsmoker, is otherwise healthy, and his physicians believe he has a high likelihood of a surgical cure.

In contrast, Mr. B had a widely metastatic adenocarcinoma of the lung. He had exhausted current aggressive therapies and desired to be kept comfortable as he neared death. He requested "comfort measures only," declining even routine blood tests. Mr. B also wanted "do not resuscitate" and "do not intubate" status: If his respirations or heart stopped, clinicians would not intervene. At home under hospice care, he was placed on round-the-clock pain medications and other drugs to control uncomfortable symptoms; most of the drugs were delivered through patches placed on his skin. With family and friends at his bedside, Mr. B died without evident pain or distress.

Both scenarios involve an adenocarcinoma of the lung. Mr. A had a high risk of incurring high costs (for extensive diagnostic workup and major surgery), but he had a low risk of imminent death. In contrast, Mr. B's care was relatively inexpensive (home-based pain control), but he had a high risk of dying soon. Mr. A received aggressive treatment aiming for a cure; Mr. B desired to maintain comfort without intensive intervention. Mr. A and Mr. B chose clinicians and institutions that best met their personal goals and clinical needs. Mr. A sought care at a major academic medical center, whereas Mr. B obtained care in his community, close to family and friends. Thus, comparisons of cost and mortality outcomes for lung cancer patients across different hospitals must account for differences in their patient mixes: Some institutions treat more Mr. As, whereas others see more Mr. Bs. Mr. As have high risks of incurring costs and low risks of dying, whereas Mr. Bs present the opposite scenario.

Existing risk adjustment methods typically assess risks differently (i.e., use different risk factors or weight the same risk factors differently) depending on the targeted outcome (i.e., clinical outcome, resource consumption, or patient-centered outcome). Some risk adjustment methods come in multiple versions, each version calibrated to predict a specific outcome. For example, some vendors of hospital-based risk adjustment methods have different versions for predicting hospital costs versus predicting in-hospital mortality. When using risk adjusters from a family of methods (with these different versions), users must choose the method designed for their specific outcome of interest. As suggested by the scenarios of Mr. A and Mr. B, risk adjusters designed to predict costs generally do less well at predicting deaths than do methods derived specifically for mortality analyses.

Over What Time Frame?

Risks must be framed within specific time windows. As an extreme example, calculating the risk of death is moot if the time window involves lifetimes; everybody faces a 100 percent risk of dying. Mr. A may not die for decades, whereas Mr. B's death occurred in a few weeks. Similarly, costs are typically measured within explicit time frames, such as a hospitalization or a year of care. Time frames thus clarify the outcome of interest and suggest which risk factors are most important. Chapter 4 addresses in detail the issues raised by different time frames.

The time frame generally determines the data sources for risk factors and vice versa. For example, as described in Chapter 5, many studies rely on computerized hospital discharge abstracts; their diagnosis and procedure codes represent the entire hospitalization, although the recent addition of "present on admission" flags may narrow this time frame. Risk adjusters modeling hospital outcomes to identify quality shortfalls aim to capture risk factors that predate care. Otherwise, serious clinical findings (potential risk factors) could become confounded or confused with substandard care. For quality assessment, the time window for extracting risk factors helps determine the *attributional validity* of the risk-adjusted outcomes information—the likelihood that poor risk-adjusted outcomes reflect poor care rather than high patient risks (see Chapter 9). The attributional validity of narrower pretreatment time windows is presumably superior to that of wider windows.

Risk adjusters predicting costs over a year, such as those intended to set capitated payment levels for managed care health plans, often have two versions, both typically derived from computerized claims or encounter records. *Concurrent* models use data from a particular year to predict costs for that same year, whereas *prospective* models predict costs for the following year. Obviously, predicting future costs is more difficult than retrospectively modeling concurrent costs. The preferred approach depends on the purpose.

Perceptions of outcomes can change substantially with even small shifts in the window of observation. An excellent example is the short-lived Cleveland Health Quality Choice (CHQC) program, a voluntary coalition of businesses, hospitals, and physicians, which involved gathering detailed clinical data from medical records to risk-adjust hospital outcomes for general medical, surgical, and obstetrical patients. Initial CHQC data were released privately to hospitals for internal improvement activities; the first public report was issued in April 1993. The program gathered data from 1991 through 1997, during which Cleveland's absolute, risk-adjusted in-hospital mortality rates declined by up to 4.8 percent (Baker et al. 2002a). Had quality of care actually improved that significantly? Probably not. When Baker and colleagues (2002a) looked instead at mortality 30 days after hospital admission (a fixed time window), the rates did not change significantly. Between 1991 and 1997, deaths had shifted from Cleveland hospitals to other settings soon after discharge.[1]

For What Population?

The US population is remarkably diverse. Many of us are exquisitely aware of this diversity on a daily basis because of the effect our visible attributes have on our interactions with others. From the earliest moments of self-recognition, we learn our age, sex, skin color, and language and perceive the immediate world around us. As we grow, our interpretations of these basic dimensions modulate, and their meanings expand. Age becomes generation; sex becomes gender (and gender identification, with all its complexities); skin color becomes race; language becomes ethnicity; and the immediate world, with its myriad complexities, becomes culture. We develop other, sometimes shifting associations with economic class, educational attainment, religion, occupational group, disability, political ideology, and other identities.

These many dimensions, alone or in combination, help delineate populations or subpopulations that have different risks for various health-related outcomes (see Chapter 3). Some distinctions are self-evident: Children, on average, face lower risks of imminent death than do persons in extreme old age. Women and men have different risks for certain diseases. As described in Chapter 3, troubling risks arise not from intrinsic individual factors (i.e., not from biological or physiological differences), but from disparities in the way people are treated in our health care system or society at large because of their characteristics.

The population of interest helps determine the range of risk factors required for assessing the specified outcome within the pertinent time frame. For example, when examining intensive care unit (ICU) mortality rates, the relevant physiological parameters vary among neonates, children, and adults, although immediate acute findings are particularly relevant. Depending on how populations are defined, some outcomes are more pertinent than others. This book contains four chapters on specific populations: children (Chapter 13), persons with mental health conditions (Chapter 14), persons with disabilities (Chapter 15), and individuals receiving long-term care in institutional and home-based settings (Chapter 16). Especially relevant outcomes vary across these populations. Although children experience similar life-and-death outcomes as adults, albeit at different rates, specific functional outcomes differ (e.g., school performance and developmental milestones for children; productive employment for working-age adults). Important outcomes for persons with psychiatric disorders emphasize mental and emotional health and ability to perform routine social roles. For persons with disabilities and long-term care populations, functional abilities and performance of daily activities are key outcomes.

The target population reflects the underlying purpose of the developers of the risk adjustment method. For example, the designers of the Pediatric Risk of Mortality Score (PRISM) were explicitly interested in children treated in ICUs (Pollack, Ruttimann, and Getson 1987), while the developers of the

Acute Physiology and Chronic Health Evaluation (APACHE) focused on adult ICU patients (Knaus et al. 1981). The purpose generally determines the population and thus typically indicates the relevant data source. The developers of the Chronic Illness and Disability Payment System (CDPS), for example, wanted to create a method of capitating payments specifically for Medicaid recipients (Kronick, Zhou, and Dreyfus 1995). By using Medicaid databases, they created a risk adjuster calibrated to impoverished persons, primarily women and their children and persons with disabilities.

For What Purpose?

Answers to the three previous questions (risk of what, over what time frame, and for what population) are driven by the purpose of risk adjustment. As described in Chapter 1, the underlying motivation of risk adjustment is comparison: contrasting outcomes or performance for individual patients, groups of patients, or populations to those of their counterparts. The following are examples of potential purposes:

- To set payment levels for individual patients (e.g., DRGs for acute care hospitalizations) or health plan enrollees (e.g., capitation payments)
- To encourage providers or health plans to treat or accept high-cost or potentially high-risk patients
- To compare efficiency and costs of care across providers or health plans
- To compare clinical or patient-centered outcomes across providers or health plans
- To produce public report cards about performance of individual providers, as on CMS's Hospital Compare website
- To compare patient outcomes across physicians or services in an individual practice or institutional setting to guide and monitor quality improvement

The purpose dictates how well the risk adjuster must perform to succeed (i.e., to produce valid comparisons; see Chapter 9). For example, methods designed to predict costs over one year rely on administrative data, which are often messy and contain limited clinical information (see Chapter 5); these risk adjusters typically explain less than 25 percent of the variation in future costs. Nevertheless, this performance is far superior to adjustments using only demographic information (e.g., age, sex), and it meets the needs of important purchasers such as Medicare and Medicaid.

Another purpose for risk adjustment is to motivate quality improvement. Without this adjustment, clinicians or institutions with poor outcomes

could argue that they are treated unfairly: "Our patients are sicker; that's why our results are worse." Most clinicians will not believe and act on results if they do not consider the risk adjuster to be clinically credible. For a risk adjuster to be perceived as such, additional data collection and in-depth review of the clinical logic underlying the risk adjustment model by the participating clinicians may be required (see Chapter 8).

No risk adjuster is perfect. As described in Chapter 3, adjusting for all patient characteristics is neither necessary nor possible. Therefore, efforts shift to identifying risk factors that are sufficiently valid for the explicit purpose. Statistical measures of model performance alone (e.g., percentage of variation explained; see Chapter 10) do not determine validity. Such measures reveal little about whether systematic errors in predictions occur for selected subpopulations or whether important risk factors are included appropriately.

For some purposes, ethical concerns raise questions about whether and how to risk-adjust. Such situations arise when persons with certain attributes (e.g., gender, race, socioeconomic status) that might be potential risk factors for a given outcome simultaneously face the likelihood of receiving substandard care because of those attributes. One example involves performance reports that compare rates of routine screening tests or preventive services for enrollees of different health plans. Outcomes (here, technically, processes of care[2]) that depend on patients' actions (e.g., having a mammogram, having an infant immunized) raise special concerns. Education, motivation, wherewithal (e.g., transportation, child care, time off from work), care and outcome preferences, cultural concerns, and a host of other factors affect whether patients take these actions. Different health plans and providers see different mixes of patients along these critical dimensions. Therefore, from a purist's perspective, risk adjustment is indicated. However, evidence suggests that racial and ethnic minorities and persons with low socioeconomic status obtain preventive services at lower rates, probably because of a complex mix of factors but also potentially as a result of discriminatory attitudes. As Romano (2000, 978) observed:

> Before instituting case-mix adjustment of health plan or provider performance measures, we must consider both the hidden assumptions and the potential consequences. One assumption is that persons of lower socioeconomic status *inherently* use preventive services less than persons of higher socioeconomic status. If culturally sensitive, readily accessible systems of care can eliminate or substantially reduce sociodemographic disparities . . . then adjusting for case mix would implicitly "excuse" health plans for failing to implement disparity-reducing innovations. . . . [Plans might also find that] it is easier to boost [their] scores by focusing on better educated, easier-to-reach members. A related implication is that we should accept lower performance, or set lower performance targets, for plans that enroll diverse populations.

Therefore, risk adjusting for these sociodemographic attributes seems inappropriate given the ultimate purpose of using outcomes data to motivate improvement for *all* patients. Risk stratification, described in Chapter 3, is a simple solution that could also yield useful insight about how different subpopulations fare.

Even when applied to the same data set, different risk adjusters can produce different answers about the outcome of interest, and it is sometimes impossible to determine which method's answers are "right" (i.e., which method produces results that best represent the underlying truth). This difficulty could complicate decisions about which risk adjuster best meets the purpose. For example, in 2006, Massachusetts established its Health Care Quality and Cost Council, aiming to set strategies for improving health care quality while controlling costs and eliminating racial and ethnic disparities in care. To motivate Massachusetts hospitals to improve their quality of care, the Council issued a proposal to develop and publicly report overall hospital mortality rates (i.e., mortality rates reflecting care across the entire hospital rather than mortality rates for specific diagnoses or procedures).

Recognizing the analytic complexity of risk adjusting and producing these hospital-level mortality rates, the Council asked the Massachusetts Division of Health Care Finance and Policy (DHCFP) to evaluate methods for this purpose. In November 2008, DHCFP issued a call for methods of producing hospital-wide mortality measures that could be used for quality improvement and public reporting. Five commercial vendors responded: 3M Health Information Systems (3M), using its All Patient Refined Diagnosis-Related Groups (APR-DRGs); the Dr. Foster Unit at Imperial College London (Dr. Foster); Thomson Reuters; University HealthSystem Consortium (UHC); and Premier. The latter two, UHC and Premier, decided to collaborate to develop a new UHC-Premier method. All vendors received identical standard abstract information on 2,528,624 discharges from Massachusetts acute care hospitals from October 1, 2004, through September 30, 2007. The vendors applied their risk adjustment algorithms to predict probabilities of in-hospital death for each discharge and hospital level observed and expected mortality rates (Shahian et al. 2010).

Despite using the same data, the four risk adjustment methods produced different results (Shahian et al. 2010). One explanation may be their use of different subsets of discharges, despite the ostensible purpose of looking at hospital-wide mortality. As shown in Exhibit 2.1, 3M considered 95 percent of the total discharges, while UHC-Premier analyzed only 28 percent of discharges. All four risk adjustment algorithms were applied to only 22 percent of the hospital discharges. The methods also considered patients with differing characteristics (Exhibit 2.1), including variations in average age and type of case (e.g., childbirth, neonates, mental health conditions). Each method calculated

Characteristic	Risk Adjustment Method				All Discharges	Discharges Used by All Four Methods
	3M	Dr. Foster	Thomson Reuters	UHC-Premier		
Hospitals						
No. included	83	82	83	81	83	81
% of total	100	99	100	98	100	98
Discharges						
No. included	2,406,881	1,072,918	2,048,377	716,315	2,528,624	567,784
Median no. per hospital	21,926	10,140	18,747	7,114	23,428	5,870
% of total	95	42	81	28	100	22
Mean length of stay (days)	5.8	6.1	5.7	7.0	5.8	6.5
Inpatient mortality rate (%)	2.0	4.0	2.4	5.9	2.1	6.2
Mean age (years)	50 (28)*	52 (33)	56 (24)	66 (23)	51 (28)	69 (20)
Selected diagnoses (%)**						
Respiratory	9.7	17.1	11.2	25.0	9.9	26.2
Circulatory	15.1	25.5	17.1	25.9	15.5	30.4
Pregnancy or childbirth	10.4	<0.1	12.2	<0.1	10.0	<0.1
Newborn or neonate	10.0	21.7	<0.1	3.3	9.8	0
Mental disorders	4.7	<0.1	5.1	0	4.6	0

EXHIBIT 2.1 Hospitals and Patient Characteristics by Risk Adjustment Method and All Hospital Discharges in Massachusetts: FY2005–FY2007

*Mean age in years (standard deviation)

**Five conditions from among the top ten major diagnostic categories that accounted for more than 5 percent of discharges included by at least one risk adjustment method

Source: Adapted from Shahian et al. (2010, 2533).

a probability of death for each discharge it analyzed. Exhibit 2.2 shows the Pearson correlation coefficients from 2007 data comparing these predicted probabilities of death between pairs of methods for the roughly 22 percent of discharges considered by all four methods.[3]

Given that a primary purpose of these data was public reporting, differences across the methods with regard to hospitals' mortality rates (higher or lower than expected)—after accounting for their populations' risk factors—were especially sobering. For each of the four risk adjustment methods and for each of the three years, Exhibit 2.3 shows the number of hospitals considered to have either higher- or lower-than-expected mortality rates. As shown, the

EXHIBIT 2.2 Correlation Coefficients for Agreement Among Discharge-Level Estimates of In-Hospital Risk of Death*

Method	Correlation Coefficient			
	3M	Dr. Foster	Thomson Reuters	UHC-Premier
3M	1.00	0.61	0.58	0.67
Dr. Foster		1.00	0.48	0.59
Thomson Reuters			1.00	0.59

*Correlation coefficients from 2007 Massachusetts hospital data
Source: Adapted from Shahian et al. (2010, 2534).

methods differed substantially in the numbers of hospitals considered to be outliers, especially in 2005. According to Shahian et al. (2010, 2534):

> Kappa statistics [see Chapter 9] indicated poor-to-substantial agreement between methods in classifying hospital mortality performance, depending on the year and method pairs. In fiscal year 2007, kappa statistics for agreement between methods in designating higher-than-expected outliers ranged from −20.04 (UHC-Premier and 3M) to 0.39 (Thomson Reuters and Dr. Foster). For some individual hospitals, categorizations varied widely and in some cases were completely discordant. For instance, in fiscal year 2006, of 28 hospitals designated as having higher-than-expected hospital-wide mortality by one method, 12 were simultaneously classified as having lower-than-expected mortality by other methods (6 by one method, 3 by two methods, and 3 by three methods).

Exhibit 2.4 shows plots, by pairs of risk adjustment methods, of hospitals' standardized mortality ratios produced by each method and multiplied by 100 (see Shahian et al. [2010, 2536] for details about computations). Especially striking were the differences between the rates produced by the Thomson Reuters method and those produced by the other vendors' methods for a facility labeled Hospital C, likely because the Thomson Reuters method used only 3 percent of Hospital C's discharges to calculate the standardized mortality ratios while the other methods used at least 30 percent. The 3 percent used by Thomson Reuters had a much larger number of high-mortality diagnoses (e.g., respiratory diseases and cancers) and a lower number of low-mortality diagnoses (e.g., childbirth) than did discharges overall. Thus, the 3 percent included in Thomson Reuters' calculations for Hospital C had a mortality rate of 59.8 percent, while Hospital C's overall mortality rate was just 2.2 percent.

Findings from this Massachusetts project prompted Shahian and colleagues (2010) to question which method's results were right or the truest reflection of hospital quality (as proxied by hospital-wide mortality rates). Given that the purpose of producing these figures was to provide insight into hospital quality, their question was central to the study, but one that they could not answer. They did not have an independent, "gold standard" indicator of

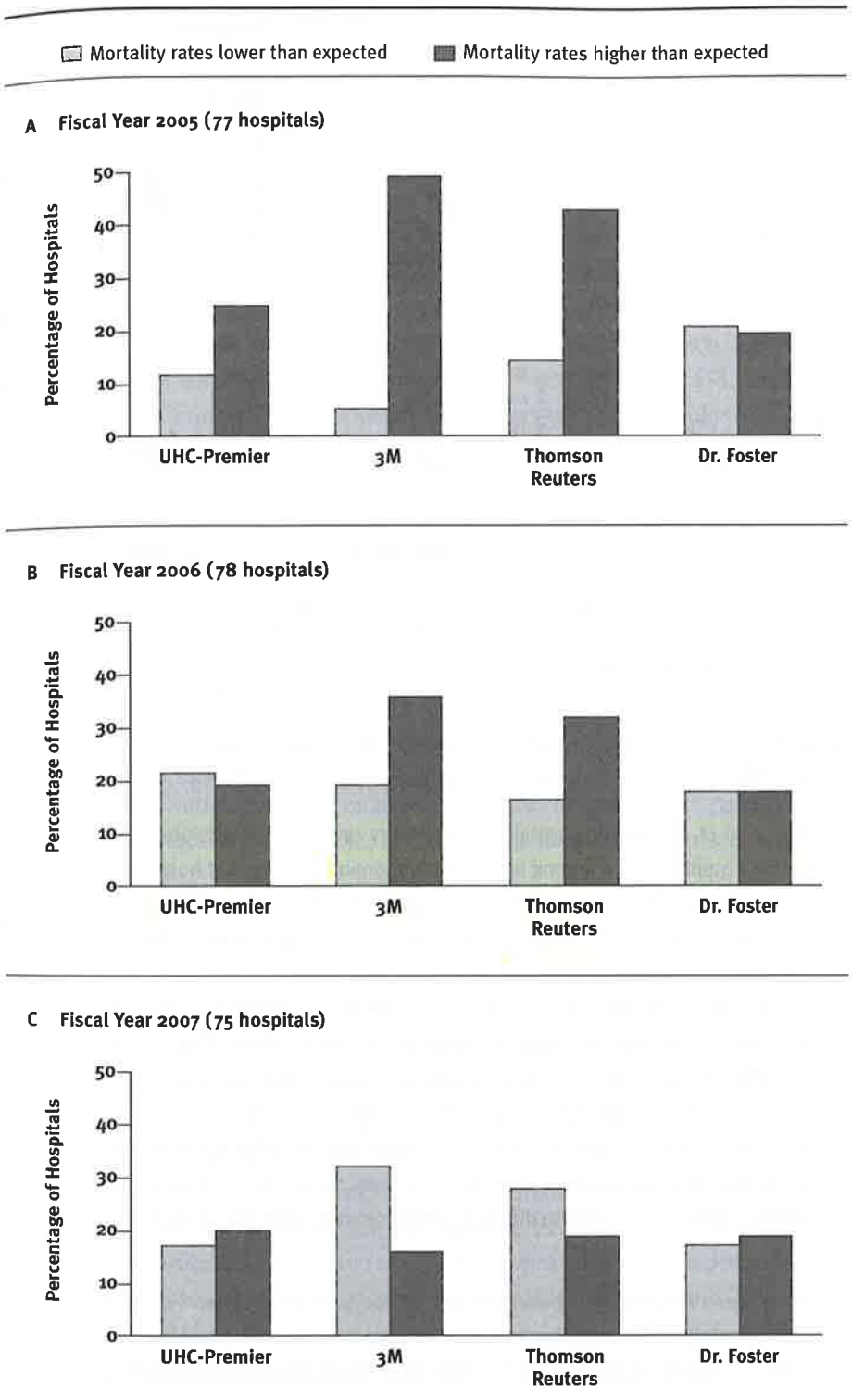
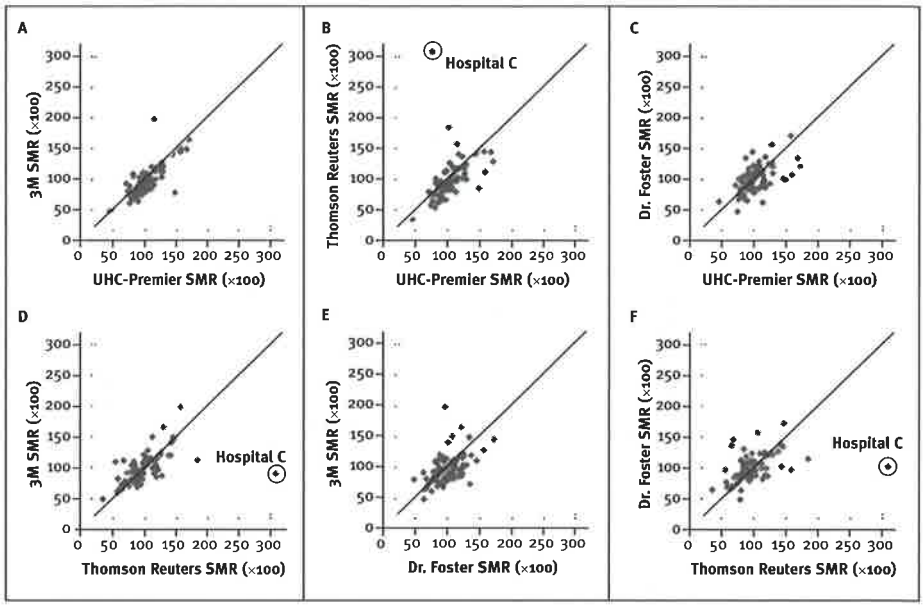


EXHIBIT 2.3 Percentages of Massachusetts Hospitals with Mortality Rates Higher or Lower than Expected According to Four Risk Adjustment Methods: FY2005–FY2007

Source: Reprinted from Shahian et al. (2010, 2537); copyright held by Massachusetts Medical Society. Used with permission.

EXHIBIT 2.4
Comparison of Hospital-Level Standardized Mortality Rates According to Four Risk Adjustment Methods: 2007



Source: Reprinted from Shahian et al. (2010, 2536); copyright held by Massachusetts Medical Society. Used with permission.

hospital quality against which to compare the findings of the risk adjustment study. The researchers concluded (Shahian et al. 2010, 2534–35):

> The goal of assessing hospital-wide mortality rates is to make inferences about the relative quality of care among hospitals. Proponents believe that hospital-wide mortality metrics provide useful warning flags about problems with the quality of inpatient care, aid consumers in choosing a hospital, and help provide a focus for hospital quality-improvement activities. . . .
>
> The four commercially available methods for assessing hospital-wide mortality that we studied are marketed to hospitals to support internal quality-improvement activities. However, their implications are even more important, and the corresponding need for methodological accuracy is greater, when such measures are used for broader initiatives, such as public reporting or performance-based purchasing. We found that estimates of hospital-wide mortality could vary, sometimes widely, among methods, which consequently leads to different inferences regarding the quality of hospital care.
>
> . . .
>
> Differences in categorizing performance on the basis of hospital-wide mortality rates raise the inevitable question of which method best identifies potential quality problems. Our study could not address that question, since an observable benchmark for overall hospital quality does not exist. . . . This disagreement suggests that all methods are not reflecting the same underlying construct, although it is possible that one method might perform better than the others in estimating the quality of hospital care.

Additional Considerations

Answering the four major questions outlined earlier is simply the beginning; many important issues remain. The most crucial practical consideration is the data source. Will the risk adjuster rely on standard, coded administrative data (see Chapter 5); clinical information from medical records, which should become increasingly electronic in coming years (see Chapter 6); or direct responses from patients (see Chapter 7)? The nature of the database shapes the design of the risk adjustment method: With large data sets, analysts can develop and test risk adjusters empirically, whereas without such data, measures must rely on clinical judgment. As described in Chapter 8, the most statistically and conceptually robust risk adjusters generally result from interactions between clinicians and statistical modeling.

The data also delimit the range of candidate risk factors (see Chapter 3). One important distinction, especially when the purpose of the analysis is to predict costs, is whether the risk adjuster considers procedure use. Methods designed to predict short-term costs (e.g., DRGs) rely heavily on procedural information, particularly the presence and type of major surgery: The costs of operations generally overwhelm the costs of medical or recuperative care. Not surprisingly, DRGs overall are poor risk adjusters for hospital mortality, although they perform slightly better within surgical DRGs (Hofer and Hayward 1996). Because the use of many procedures is highly discretionary, risk adjusters targeting clinical outcomes generally eschew procedures in rating risk. As pharmacy data become increasingly available, some risk adjusters are using them to proxy disease burden.

Finally, selection of appropriate and reasonable analytic techniques raises important questions. As described in Chapter 11, risk adjustment is often used to examine the results of observational studies in which patients are not randomly assigned to various treatments or care plans. In many cases, sample sizes are small and there is no single right way to analyze the data. Nevertheless, analytic choices can carry important implications, as suggested in chapters 10 through 12. Consideration of these implications is essential to meaningful interpretation of risk-adjusted outcomes information.

Even after risk adjustment, questions remain about what comparative outcomes information really means. For example, even with optimal risk adjustment, do risk-adjusted mortality rates provide meaningful clues about hospital quality (Pitches, Mohammed, and Lilford 2007; Lilford and Pronovost 2010; Black 2010)? Answering this question in a meaningful fashion may prove even more vexing than designing the risk adjustment methodology. Nonetheless, as Nightingale and Codman said (see Chapter 1), this question about quality highlights the ultimate purpose of gathering and analyzing the data—to motivate and guide improvements.

Notes

1. CHQC disbanded in July 1999 because hospitals affiliated with Cleveland Clinic refused to voluntarily submit their data. Since the outset, Cleveland Clinic officials complained that the CHQC risk adjustment ignored special characteristics of their patient population (Vogel and Topol 1996). They also protested against paying $2 million annually for data collection. As further justification for its withdrawal, the Cleveland Clinic noted that the CHQC data were not being used.
2. Performance measures are often sorted into two types: outcomes and processes. The rationale for risk adjustment of outcomes (i.e., how patients fare) is generally clear. Process measures (i.e., what is done for patients) may also warrant risk adjustment, although the conceptualization of the measures dictates the extent of adjustment required. Some observers increasingly blur the semantics distinguishing outcome from process measures. For example, is obtaining a mammogram an outcome or a process of care? Many quintessential process measures build in explicit information about patient characteristics that are essentially risk factors for obtaining the service. For example, the use of beta-blockers after a heart attack is a widely accepted process measure, but with the stipulation that patients not have certain contraindications to receiving the beta-blockers. Risk adjustment for comparing beta-blocker use across providers does not need to control for contraindications to those drugs because persons with those conditions are eliminated from consideration.
3. Exhibit 2.2 shows Pearson correlation coefficients derived from 2007 data. Figures from 2005 and 2006 data were comparable (Shahian et al. 2010, 2534).

CHAPTER

3

RANGE OF RISK FACTORS

Lisa I. Iezzoni

Complex, multidimensional factors affect the health and well-being of humans. We function on numerous levels within multiple spheres: as living organisms, our biology largely (although not exclusively) determined by intricate, molecular road maps of genetic codes; as thinkers, actors, and communicators who shape and are shaped by our social and physical environments; as members of families, communities, cultures, and populations; and as individuals with different goals, behaviors, preferences, and expectations about our lives and futures. From the physiological to the existential, these dimensions carry consequences for our individual health and wellness, some directly and others more circuitously. Given these effects, each dimension could theoretically serve as person-specific risk factors in some über risk adjustment methodology.

No risk adjustment method, however, can account for all relevant risk factors—at least not yet. Data on all dimensions of risk simply do not exist. Nevertheless, the most important attribute of any risk adjustment approach is the set of risk factors that it does consider. The scope of risk factors largely determines whether risk-adjusted results are credible and valid (see chapters 8 and 9). The risk factors dictate whether a method uses sufficient information to convey clinical expectations about patients (i.e., is medically meaningful). Given that in many cases the ultimate goal of the risk-adjusted information is to affect clinicians' behavior or gain their cooperation, medical meaningfulness is essential.

As described in chapters 5 through 7, data limitations impose immutable constraints on risk adjustment methods. Collection of information on all potential risk factors is logistically and practically infeasible. Because risk adjustment methods cannot consider all risk factors, how credible and trustworthy are the findings when risk adjustment Method A is used to answer Question X? To answer this question we need an *a priori* conceptual model delineating which risk factors a risk adjustment method should consider for a given outcome, time window, population, and purpose. With that conceptual model, analysts can determine which risk factors are and are not addressed by Method A. Understanding which risk factors are conceptually important but excluded suggests how cautiously or confidently one should interpret the risk-adjusted results for Question X. Are the findings believable?

Do differences in outcomes result from unmeasured risk factors or from variations in therapeutic effectiveness, quality of care, efficiency, or another aspect of care?

This chapter enumerates and describes human attributes that could be important risk factors in specific settings. Exhibit 3.1 lists potential risk factors by broad category, and Exhibit 3.2 presents 13 brief clinical synopses reflecting these risk factors. Implied distinctions among these attributes are artificial. Concepts often overlap and are seldom clinically separable, especially for individual patients. I discuss intertwined characteristics individually to review them in an organized and systematic fashion while recognizing that the relevance of particular risk factors varies by outcome, time window, population, and the purpose of risk adjustment. The risk factors appearing in Exhibit 3.1 are not a comprehensive list covering all possible indicators of risk; furthermore, researchers are constantly identifying new relationships and risk factors (e.g., biomarkers, specific genes) that could become prominent in the future. Chapter 4 considers how the window of observation affects the role of various risk factors, and Chapter 8 describes how researchers might choose among various factors in specifying a risk adjustment model.

Genetics and Genomics

Viewed simplistically, genes—deoxyribonucleic acid (DNA) or genetic makeup—represent the ultimate in risk factors. DNA largely determines the biological nature of virtually every living organism on earth. Structured as a double helix, DNA is composed of two linked strands that twist around each other. Each strand contains four different types of molecules called *nucleotide bases*: adenine (A), thymine (T), guanine (G), and cytosine (C). Functioning like a genetic alphabet, each base links with only one other type of base in its opposite strand: A always pairs with T, and C with G. Human DNA contains approximately three billion nucleotide base pairs (National Human Genome Research Institute 2011). Sets of these nucleotide base pairs compose genes, which typically carry instructions for producing a particular protein or set of proteins. The human genome contains about 20,000 to 25,000 genes, arrayed along 23 pairs of chromosomes (46 total chromosomes) within the nucleus of each cell in the body.[1] As cells divide and multiply, chromosomes from parent cells replicate and populate the nucleus of each progeny cell. Sometimes mistakes or mutations occur during this replication process, giving the new cells different instructions for protein production or other molecular activities and thereby allowing specific abnormalities (e.g., diseases) as well as potentially beneficial longer-term changes (e.g., evolution) to occur.

Future risk prediction models for human diseases and even human behaviors might include detailed information about the genes of each individual

EXHIBIT 3.1
Range of Risk Factors for Health and Well-Being Outcomes

Genetics
- Genetically determined traits
- Genetic predisposition to specific health conditions
- Genetic predisposition to specific health-related behaviors

Demographic characteristics
- Age
- Sex
- Race and ethnicity
- Primary language, level of English proficiency
- Country of origin and immigration status

Clinical factors
- Acute physiological stability
- Diagnoses and other health conditions
- Extent and severity of diagnoses
- Sensory functioning (vision, hearing)
- Physical functional status
- Cognitive status
- Mental and emotional health

Psychosocial, socioeconomic, and environmental factors
- Familial characteristics and household composition
- Educational attainment, health literacy
- Marital or partner status
- Sexual orientation
- Domestic violence
- Economic resources
- Employment and occupation
- Housing
- Neighborhood characteristics
- Urban, rural residence, geographic region
- Health insurance coverage, underinsurance

Health-related behaviors and activities*
- Tobacco use
- Diet and nutrition
- Obesity and overweight
- Physical activity
- Sleep
- Excessive alcohol use
- Illicit drug use
- Unsafe sexual practices

Quality of life, attitudes, and perceptions
- Overall health status and quality of life
- Cultural beliefs and behaviors
- Religious beliefs and behaviors
- Preferences and expectations for health care services
- Do not resuscitate (DNR), "comfort measures only" choices for end-of-life care

*Some risk factors listed under this heading also may be affected by genetic makeup, underlying physiology, or external factors or processes rather than being solely behavioral (i.e., related to individuals' personal choices).

EXHIBIT 3.2
Thirteen Clinical Synopses: Adenocarcinoma of the Colon

Patient	Clinical Synopses
A	Family history of familial adenomatous polyposis (FAP). Genetic testing shows Adenomatous Polyposis Coli (APC) gene on chromosome 5q21. Periodic surveillance colonoscopies identify numerous polyps when patient is 45 years old. Patient chooses total colectomy to eliminate risks of adenocarcinoma of the colon.*
B	Family history of FAP. Genetic testing shows APC gene on chromosome 5q21. Periodic surveillance colonoscopies identify numerous polyps when patient is 45 years old. Patient chooses rectum-sparing partial colectomy to reduce adenocarcinoma risk. Requires ongoing surveillance sigmoidoscopy to monitor for rectal adenocarcinomas.
C	No symptoms, no family history. Microscopic nidus of well-differentiated adenocarcinoma found in polyp during a routine screening colonoscopy. Considered a surgical cure. Otherwise in good health.
D	No symptoms. Microscopic nidus of well-differentiated adenocarcinoma found in polyp during a routine screening colonoscopy. Considered a surgical cure. History of stroke one year ago. Partially paralyzed on right side of body; has difficulty speaking.
E	No symptoms. Microscopic nidus of well-differentiated adenocarcinoma found in polyp during a routine screening colonoscopy. Considered a surgical cure. Patient refused medication to treat serious depression; recently lost job. No known family or friends.
F	Patient septic with *Clostridium perfringens* bacteria. Polyp found on colonoscopy; microscopic nidus of well-differentiated adenocarcinoma found in polyp. Otherwise in good health.
G	Large adenocarcinoma with invasion deep into wall of colon. No evidence of distant metastases. Otherwise in good health.
H	Large adenocarcinoma with invasion deep into wall of colon. No evidence of distant metastases. History of poorly controlled essential hypertension, with routine blood pressure readings of 164/94 mmHg.
I	Large adenocarcinoma causing bowel obstruction. No evidence of distant metastases. Patient acutely septic and in shock.
J	Widespread metastatic disease. Patient desires active intervention, including ICU admission and intubation if necessary.
K	Widespread metastatic disease. Patient requests DNR status, desiring comfort measures only.
L	Patient doing well clinically but dies unexpectedly.
M	Patient doing poorly clinically and dies as expected.

*Persons with FAP also have much higher risks of other types of cancer, including tumors in the thyroid, small bowel, liver, and brain, than do persons without this condition.

in the population of interest. That day is not yet here, although our understanding of human genes and their contributions to health and disease has grown exponentially over the last decade. In April 2003, the Human Genome Project finished mapping the sequence of human genes, using cells taken from the blood of selected volunteers. Today, efforts continue to produce more comprehensive maps of genetic variability across individuals, with the goal of mapping the functions of all human genes, their roles in diseases, and consequences for human–environmental interactions. This mapping has enabled investigators to move from genetics (i.e., the study of single genes) to genomics (the study of functions and interactions among all genes in the genome) (Guttmacher and Collins 2002). Huge databases containing genetic information from many individuals, such the International Human HapMap Project, are producing insights into the genomic risks associated with such common conditions as type 2 diabetes and rheumatoid arthritis (Hardy and Singleton 2009). With more genomic data comes an increasing understanding of the complexity of factors affecting gene expression.

Genes have varying and complicated roles as risk factors for disease (Rotimi and Jorde 2010). Sometimes a single gene largely determines whether a disease will occur, while other times multiple or interacting genes increase the risk of developing a disease. Colon cancer, the third most commonly diagnosed cancer in men and women in the United States, is an example of a disease that may develop out of either scenario (effects of a single gene versus multiple or interacting genes). Because of a genetic mutation inherited from one parent—the Adenomatous Polyposis Coli (APC) gene on chromosome 5q21—patients A and B (Exhibit 3.2) have Familial Adenomatous Polyposis (FAP). This condition, in which more than 100 adenomatous polyps typically develop in the colon and rectum after age 10, conveys a 90 percent absolute risk of developing colorectal cancer by age 45 (National Cancer Institute 2010). FAP is genetically autosomal dominant, meaning that offspring of persons carrying the APC gene have a 50 percent chance of inheriting the disease. Fortunately, FAP is rare. Only 5 to 10 percent of colorectal cancers are related to genetic syndromes, such as FAP, that are genetically autosomal dominant (Lynch and de la Chapelle 2003). The more common scenario is for persons to have a genetic predisposition to developing cancer, as indicated by family history (Exhibit 3.3). For example, a variety of genes are implicated in the predisposition to colon cancer (Markowitz and Bertagnolli 2009).

Thus, genetics can powerfully predict individuals' future health and potential health risks. Information on genetic risks could help individuals make critical informed decisions, such as the decisions of patients A and B (Exhibit 3.2) to undergo surgery to reduce their risks of developing colon cancers. In response to concerns that health insurers or employers might use genetic information to discriminate against persons with genetic

EXHIBIT 3.3 Estimated Relative and Absolute Risks of Developing Colorectal Cancer (CRC)

Family History	Relative Risk of CRC*	Absolute Risk of CRC by Age 70
No family history	1.0	4%
One first-degree relative with CRC	2.3 (95% CI, 2.0–2.5)	9%
More than one first-degree relative with CRC	4.3 (95% CI, 3.0–6.1)	16%
One affected first-degree relative diagnosed with CRC before age 45	3.9 (95% CI, 2.4–6.2)	15%

*CI = confidence interval (statistical reliability)
Source: Adapted from National Cancer Institute (2010). Data from the Surveillance, Epidemiology, and End Results database.

predispositions to disease, Congress passed the Genetic Information Nondiscrimination Act of 2008 (GINA, P.L. 110-233 122 Stat. 881). Signed into law by President George W. Bush on May 21, 2008, GINA prohibits discrimination by employers or health insurers against individuals based on genetic information (e.g., results from genetic testing of relatives; family histories of diseases). While GINA protects genetic information, it does not cover other clinical test results, such as cholesterol levels. Furthermore, GINA's nondiscrimination provisions do not encompass life, disability, or long-term care insurance. Therefore, individuals may be reluctant to allow their genetic information to be recorded.

Almost 90 percent of studies looking at human genetic variations and associations with disease have involved populations with European ancestry (Rotimi and Jorde 2010). Recent studies are looking at large numbers of African Americans and analyzing patterns to improve our understanding of the origins of human populations around the globe and different diseases (Hinch et al. 2011). Going forward, it is important that genomic research include racially and ethnically diverse populations from around the nation and world. While information about the contributions of genes to variations in human disease is increasing, inclusion of genetic information in risk adjustment models is still premature. Studies that have added genetic information to risk prediction models to identify diseases in individual patients have yielded mixed results. Inclusion of genetic information does not greatly improve the ability of all models to predict diseases (Evans, Visscher, and Wray 2009; Kraft and Hunter 2009; Paynter et al. 2010; Pepe, Gu, and Morris 2010), but its usefulness for this purpose could change as more

populations are studied and more specific genetic associations with disease risk are identified.

In the end, genes do not determine every trait. Environmental and developmental influences often determine whether many genes are expressed (i.e., become phenotypes [attributes of living organisms]). Exhibit 3.4 demonstrates this truth for the world's first cloned cat, named CC (Shin et al. 2002).[2] Rainbow, who donated the cell nucleus containing the DNA

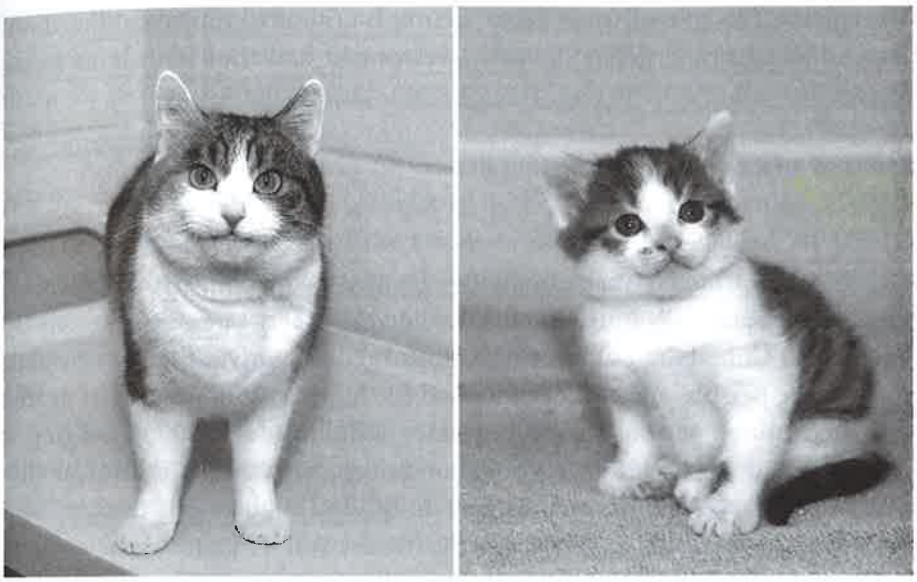


EXHIBIT 3.4A Rainbow (Donor) and CC (Clone) as a Kitten

Source: Photo used with permission from Texas A&M University College of Veterinary Medicine & Biomedical Sciences.

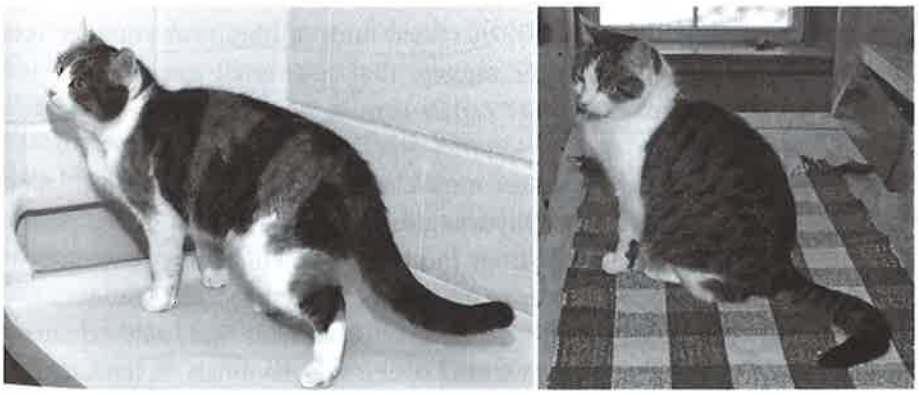


EXHIBIT 3.4B Rainbow (Donor) and CC (Clone) as an Adult Cat

Source: Photo used with permission from Texas A&M University College of Veterinary Medicine & Biomedical Sciences.

that produced CC, was a calico (tricolored) cat, but even as a kitten (Exhibit 3.4A), CC did not look like her nuclear-donor cat. The difference in appearance is even more striking in the comparison between CC fully grown and Rainbow (Exhibit 3.4B) (Shin et al. 2002, 859):

> As with other genetically identical animals with multicoloured coats, the cloned kitten's colour patterning is not exactly the same as that of the nuclear donor—this is because the pattern of pigmentation in multicoloured animals is the result not only of genetic factors but also of developmental factors that are not controlled by genotype.

Untangling the role of these other factors is critical to understanding how genetic information affects disease development and thus its role as a risk factor.

Age

Age is a characteristic that cannot be changed or manipulated. Epidemiologic studies generally must account for age. Mortality rates differ widely by age; in the United States in 2007, persons aged 5 to 9 had the lowest death rates, while persons aged 85 or older had the highest: 29.5 percent of deaths occurred among persons aged 85 or older (Miniño et al. 2009, 2). Aging is closely linked to significant chronic conditions, such as cardiovascular disease, diabetes, certain cancers, osteoarthritis, and Alzheimer's disease. Surveys of older persons, largely from the World War II generation, suggest that rates of serious functional limitations have declined significantly over recent years (Manton, Corder, and Stallard 1997; Manton and Gu 2001; Freedman, Martin, and Schoeni 2002).[3] However, studies have found higher rates of some conditions and functional impairments among the baby boom generation than among their parents' generation, raising questions about whether those trends will continue (Leveille, Wee, and Iezzoni 2005; Martin et al. 2007, 2009). Increased rates of obesity at younger ages (described later in this chapter) suggest that functional impairments and chronic conditions may occur at earlier ages in coming decades (Manton 2008).

Major scientific advances are unraveling the aging process, elucidating how aging affects tissues and physiological functioning. Instead of passive decline and deterioration, regulatory mechanisms mediated by environmental or physiological signals appear to govern the aging process. First identified in studies of short-lived organisms, such as yeast, fruit flies, and nematode worms, these regulatory pathways also operate in mammals. Scientists now raise the possibility of slowing these aging processes and thus postponing aging-related chronic conditions (Kenyon 2010; Vaupel 2010). Slowing the

biology of aging is particularly imperative to delaying or minimizing cognitive declines; today nearly half of persons aged 85 or older experience Alzheimer's disease (Bishop, Lu, and Yankner 2010).

Despite these hopes for the future, on average, older persons have worse clinical outcomes than younger persons. Symptoms and signs of disease may differ between older and younger patients. For example, as age increases, pneumonia patients become less likely to report non-respiratory and respiratory symptoms, despite having active infections (Metlay et al. 1997). Costs of care for older persons may exceed costs for younger patients because of prolonged recuperative periods and greater incidence of complications. On the other hand, because elderly patients frequently receive less aggressive treatment by personal choice, some costs of care may be lower than those for younger adults.

Age may have an effect on overall patient risk independent of other risk factors. For gravely ill patients in ICUs, age independently predicts imminent death regardless of the extent of organ system failure. APACHE III, designed to predict in-hospital mortality for ICU patients, assigns separate points for age to its score, ranging from 0 points for those younger than age 45 to 24 points for persons at least 85 years old (Knaus et al. 1991, 1624).

Depending on the clinical setting, other dimensions of risk should be viewed within the context of patients' age. One obvious distinction is among newborns, children, and adults. The Pediatric Risk of Mortality score (PRISM) uses 14 physiological variables to predict in-hospital mortality for pediatric ICU patients. In assigning points, PRISM employs different ranges for systolic blood pressure, heart rate, and respiratory rate, depending on whether the patients are infants or older children (Pollack, Ruttimann, and Getson 1988). Calculations of the probability of in-hospital mortality include not only PRISM scores but also patient age in months. Thus, PRISM weights certain physiological risk factors differently by patient age but also treats age as an independent predictor of death. Neonatal populations present special challenges to capturing age effects (Richardson, Tarnow-Mordi, and Escobar 1998). Developers of the Score for Neonatal Acute Physiology, Perinatal Extension, Version II (SNAPPE-II) wanted to use gestational age rather than birth weight as a risk factor for neonatal ICU mortality; they viewed gestational age as a more clinically valid indicator of neonatal physiology (Richardson et al. 2001). However, at the time, difficulties collecting accurate gestational age data forced them to use birth weight instead.

Standard statistical measures that quantify how well age explains outcomes, such as c and R^2 values (see Chapter 10), can be unimpressive. In a study of 30-day post-admission mortality for elderly Medicare patients, age and sex explained only 1 to 3 percent of variation in deaths (Keeler et al. 1990,

1967). In another study, age and sex produced modest c-statistics for predicting in-hospital mortality: 0.69 for acute myocardial infarction (Iezzoni et al. 1996a), 0.67 for pneumonia (Iezzoni et al. 1996b), and 0.60 for stroke (Iezzoni et al. 1995a). Among gravely ill patients, age did not substantially improve the accuracy of predictions of six-month survival after adjusting for illness severity and baseline functional status (Hamel et al. 1999a).

The importance of age may depend on the age ranges within the study populations and the outcome of interest. When examining adults defined as persons aged 18 or older, separating patients into younger and older strata may offer insight. For instance, in a study comparing illness severity among roughly 4,500 adult inpatients at teaching and nonteaching hospitals in metropolitan Boston (Iezzoni et al. 1990), severity patterns by hospital teaching status varied by age strata across different disease groups. Only younger patients with coronary artery disease and younger patients with low back pain appeared sicker at teaching hospitals than at nonteaching hospitals; heart attack severity was higher at teaching facilities among both younger and older patients. The researchers hypothesized that patients' preferences and physicians' practice patterns may cause certain differences. Younger and older patients may vary in their willingness to travel to distant, inner-city hospitals or to seek intensive, high-technology care.

Increasing numbers are joining the oldest old, variably defined as older than age 80 or 85. Scientists are examining the biological mechanisms that allow some people to survive well beyond average life expectancies. Studies are exploring such factors as the length of telomeres, structures located at the ends of linear chromosomes that protect against DNA degradation. Exceptionally healthy oldest-old individuals ("super seniors") appear to have different telomere lengths than other persons (Goronzy, Fujii, and Weyand 2006; Halaschek-Wiener et al. 2008). Persons who are exceedingly old, notably centenarians, may also have different genetic patterns (Sebastiani et al. 2010). Regardless of the genetic mechanisms, oldest-old patients have lower physiological reserves (i.e., lower ability to rebound from the physical assaults of acute illness) than younger patients have and are more likely to develop complications. Despite these inevitabilities, carefully selected oldest-old patients can benefit significantly from major interventions and experience improved post-procedural functional status and quality of life, although the research evidence base for this population is often limited (Alexander et al. 2007; Jokhadar and Wenger 2009). In a study of more than 30,000 heart attack patients, persons aged 85 or older were significantly more likely than younger patients to be women, to have hypertension, and to experience heart failure or stroke (Forman et al. 2010). However, only 10 percent of persons aged 85 or older had absolute or relative contraindications to reperfusion procedures for these conditions.

The primary explanation for not receiving reperfusion interventions was patients' preferences (45 percent).

Nonetheless, when adjusting for age in observational studies, the possibility of ageism merits consideration. Increasing numbers of elderly patients undergo sophisticated invasive treatments. However, older patients may receive fewer screening services than younger patients (Schonberg, Leveille, and Marcantonio 2008), although these findings vary by type of screening service (Schonberg et al. 2008) and some screening services may not be indicated beyond certain ages. Treatments for the oldest old may also be less intensive than for younger individuals with similar clinical presentations and preferences (Hamel et al. 1999b, 2000). For example, early-stage breast cancer patients aged 80 or older receive considerably less aggressive care than younger women and are more likely to die from their cancers (Schonberg et al. 2010). However, ageism should be disentangled from varying preferences among older patients; one study found that although elderly cancer patients received less aggressive care, they spent more time discussing limitations of treatments with their physicians (Rose et al. 2000). Older patients may have different expectations of health care services and are often more satisfied than younger patients with the outcomes of the services they receive. Among major elective surgery patients in one study, elderly persons had similar global health perceptions as did younger patients despite worse physical and role function, lower energy, and greater fatigue (Mangione et al. 1993).

Age is easy to understand and explain, with good face validity (see Chapter 9) as an important risk factor for many different outcomes. Information on age is generally readily available and reliable; while often statistically insignificant, age is therefore standard in risk adjustment models.

Sex and Gender

Women and men differ chromosomally, anatomically, physiologically, and hormonally. Few of these sex-based differences are "understood in molecular or cellular terms" (Federman 2006). Nonetheless, males and females clearly face divergent risks for certain diseases and death by age strata. Women's average life spans are longer than men's; in the United States in 2007, the age-adjusted death rate per 100,000 population was 905.8 for men and 643.4 for women (Miniño et al. 2009, 1). Causes of death also vary between men and women. In the United States in 2007, the major causes of death for the total population (males and females together) across all ages were heart diseases (25.4 percent) and malignant neoplasms (23.2 percent). However, the male-to-female age-adjusted death rate ratios were 1.5 for heart disease and 1.4 for

malignant neoplasms. Males had strikingly higher age-adjusted death rates than females had for certain causes of death, with male-to-female ratios of 2.1 for accidents (unintentional injuries), 2.2 for chronic liver disease and cirrhosis and for Parkinson's disease, and 3.9 for intentional self-harm (suicide). Among the 15 leading causes of death in the United States in 2007, the only one with higher rates among women was Alzheimer's disease, with a male-to-female age-adjusted death rate ratio of 0.7 (Xu et al. 2010). Controlling for sex is therefore crucial in epidemiological studies of long-term outcomes.

As suggested by some of the cause-of-death findings (e.g., relating to injuries and suicides), differences between men and women transcend biological considerations (Gesensway 2001, 936):

> Research now shows that sex matters beyond the hormonal, anatomic, physiologic, and reproductive differences . . . . Socioeconomic circumstances that influence women's lives differently from men's—such as parenthood, poverty, or violence—can lead women to develop or react to diseases and treatments differently from men. Well-documented differences in the communication and human interaction styles of men and women affect how each sex uses the health care system, describes illnesses, or participates in decision-making with a physician.

The distinction between sex and gender is emblematic. *Sex* typically refers to the chromosomally mediated, biological distinctions between males and females, whereas *gender* relates to social roles. Distinctions become more complex when considering gender identity (self-identification of gender independent of chromosomal or physical traits) and transgender individuals, whose self-assigned gender identity does not match their chromosomal or biological sex.

The historical exclusion of women from efficacy trials of new treatments has hindered development of *a priori* hypotheses about sex as a clinical risk factor.[4] The National Institutes of Health Revitalization Act of 1993 aimed to remedy this problem by requiring inclusion of women and racial and ethnic minorities in clinical research as methodologically appropriate. However, reviews of published research suggest that women remain underrepresented in clinical research (Vidaver et al. 2000; Ramasubbu, Gurm, and Litaker 2001; Geller, Adams, and Carnes 2006). In a study of 69 randomized controlled trials published in influential journals in 2004, on average women made up 37 percent of study subjects overall and only 24 percent of subjects in drug trials; 87 percent of the publications did not present findings separately for males and females (Geller, Adams, and Carnes 2006). More recent data suggest greater inclusion of sex-specific analyses in research publications, but these statistics varied substantially across disciplines (Oertelt-Prigione et al. 2010). Furthermore, a study of 77 research articles that together made 432 claims about differences by sex found that documentation was sufficient for only 12.7 percent of those claims; most of the documentation was inadequate or spurious (Patsopoulos, Tatsioni, and Ioannidis 2007).

Sex predicts many short-term outcomes only modestly. Neither APACHE III (Knaus et al. 1991) nor PRISM (Pollack, Ruttimann, and Getson 1988) includes sex in its predictive model. A large study of Medicare beneficiaries found that sex was a significant predictor of 30-day post-hospitalization mortality for only one of five conditions: Among patients who had a hip fracture, men were at higher risk of death than women (Keeler et al. 1990). Meta-analyses of 24 studies from 1979 through 2009 confirm greater levels of higher-than-expected mortality among older men than among older women following hip fractures (Haentjens et al. 2010). A study of almost 90,000 patients admitted for one of six common medical diagnoses adjusted for clinical risk factors abstracted from medical records, including do not resuscitate (DNR) status. For four of the diagnoses, men were significantly more likely to die in hospital than women. For the other two diagnoses, men and women were equally likely to die (Gordon and Rosenthal 1999).[5]

One challenge to using sex as a risk factor for assessing patients' outcomes is the concern that men and women receive different quality of care. In the 1990s, research raised questions about disparities in treatment between men and women, especially in cardiovascular interventions. Numerous studies suggested that women with heart disease received fewer invasive diagnostic and therapeutic procedures and had higher in-hospital death or complication rates than did men (Ayanian and Epstein 1991; Krumholz et al. 1992; Watanabe, Maynard, and Ritchie 2001; Correa-de-Araujo et al. 2006). In contrast, other studies found no significant differences between male and female cardiac patients in the number of coronary artery procedures or clinical outcomes after controlling for important risk factors (Mark et al. 1994; Hannan et al. 1999). A study that used detailed clinical data from 136,247 patients suggests that higher mortality among women with acute cardiovascular disease is primarily attributable to clinical (pathophysiologic) differences in the disease between the sexes (Berger et al. 2009).

Like age, sex is a simple, routinely available variable with reasonable face validity as a risk factor in certain settings. Nonetheless, suspicions about gender bias in treatment decisions could confound risk-adjusted outcome assessments and need to be explored further. Stratification of analyses (i.e., separate examination of risk-adjusted results for men and women) could allay these concerns.

Race and Ethnicity

Several parallels with sex arise when viewing race and ethnicity as risk factors. Well-documented differences by race and ethnicity exist in longevity, disease prevalence, and leading causes of death. In the United States in 2007, age-adjusted

death rates per 100,000 population varied by broad racial and ethnic categories as follows: 961.9 for non-Hispanic black; 766.5 for non-Hispanic white; 625.3 for American Indian or Alaska Native; 530.7 for Hispanic; and 409.7 for Asian or Pacific Islander populations (Miniño et al. 2009, 4). Among the 15 leading causes of death in the United States in 2007, the ratio of age-adjusted death rates for whites compared with different racial and ethnic groups varied, sometimes dramatically. For the two leading causes—heart diseases and malignancies—the ratios of age-adjusted death rates for black to white persons were 1.3 and 1.2, respectively; in contrast, comparable ratios for Asians or Pacific Islanders to whites were 0.5 and 0.6, and for Hispanic to non-Hispanic whites, 0.7 and 0.6 (Xu et al. 2010, 5). Most notably, the ratio of age-adjusted death rates for blacks to whites for homicide was 5.7, compared with 0.6 for Asians or Pacific Islanders to whites. Social rather than genetic or biological forces drive some of these patterns.

The historical exclusion of minorities from therapeutic trials hampers the formation of *a priori* hypotheses about the clinical effects of racial or ethnic differences. Unresolved ethical and societal questions raised by the notorious Tuskegee syphilis study may have chilled efforts to recruit black persons as experimental subjects (Corbie-Smith et al. 1999), although today's younger black populations may be less aware of and thus less influenced by Tuskegee (Katz et al. 2009). Some studies suggest that black individuals are much less likely than whites to trust the research process, which may explain much of their unwillingness to participate in studies (Shaw et al. 2009; Bussey-Jones et al. 2010). Similarly, there is concern that revelations in October 2010 that the National Institutes of Health (NIH) funded a 1946–1948 study to infect nearly 700 Guatemalans with syphilis and then study their response to penicillin will dampen recruitment of Hispanic research subjects. Nevertheless, the evidence about whether minorities are less likely to join clinical trials is inconclusive. Despite passage of the National Institutes of Health Revitalization Act of 1993, a review of heart failure clinical trials from 1985 through 1999 showed no increase in the involvement of black men and women (Heiat, Gross, and Krumholz 2002). Another report examined consent rates by race and ethnicity across 20 health research studies and found little evidence of systematic differences (Wendler et al. 2006).

Race and ethnicity are hard to capture in a clear, consistent, and meaningful fashion. Researchers' definitions and operational constructs of race and ethnicity vary considerably (LaVeist 1994; Williams 1999; Comstock, Castillo, and Lindsay 2004; Hunt and Megyesi 2008; Corbie-Smith et al. 2008).

With the unraveling of the human genome, scientists have examined how genetic differences vary by race. These analyses distinguish two concepts: ancestry (genetic relationships across populations primarily defined by geography) and race (socially defined boundaries across populations). While race reflects aspects of biological ancestry, it does not equal ancestry (Bamshad

2005). Compared to many other species, humans are genetically homogenous: Only about 0.1 percent of DNA varies across humans, and this variation is most closely linked to geography (Jorde and Wooding 2004, S32):

> Because traditional concepts of race are in turn correlated with geography, it is inaccurate to state that race is "biologically meaningless." On the other hand, because they have been only partially isolated, human populations are seldom demarcated by precise genetic boundaries. Substantial overlap can therefore occur between populations, invalidating the concept that populations (or races) are discrete types.

The US Census Bureau has collected data on race since 1790. The first decennial census counted three racial categories: "whites, blacks as three-fifths of a person and only those Indians who paid taxes" (Williams 1999, 121).[6] Since 1977, Statistical Policy Directive No. 15, promulgated by the Office of Management and Budget (OMB 1994), has set federal standards for gathering racial and ethnic data. Extensive study following the 1990 census examined strategies to allow persons to self-identify more than one race and led the OMB to significantly revise the directive. Beginning with the 2000 census, Directive No. 15 stipulates that separate questions be asked about race and ethnicity. Data on race include at least five categories: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White. If persons wish to self-identify more than one race, they may do so, but a single multiracial category is not included. Ethnicity is captured through two categories: Hispanic or Latino and Not Hispanic or Latino. All federal programs implemented OMB Directive No. 15 standards by January 1, 2003.

Important practical considerations impede the collection of consistent and accurate data on race and ethnicity (Mays et al. 2003; Ford and Kelly 2005; Sohn et al. 2006). Many survey respondents misunderstand the intended distinction between race and ethnicity or feel that a dichotomous response about Hispanic origin ignores important cultural distinctions among ancestral countries (e.g., Cuba, Mexico, Puerto Rico, Argentina). Furthermore, people's responses to questions about race and ethnicity change over time or across surveys because of "fuzzy group boundaries" (ambiguities about what constitutes group membership) and "shifting identity" (Hahn 1992). Shifting identification prompted by political and social trends partially caused a 72 percent increase in the reported American Indian population between the 1970 and 1980 censuses (McKenney and Bennett 1994, 22).

The accuracy of racial or ethnic identification depends on the data source: self-reports versus external observers (Williams 1996; Williams 1999). Hospital discharge abstracts (see Chapter 5) contain racial information from patients or completely unrelated persons, such as intake receptionists or admitting clerks (LaVeist 1994). Before 2003, Veterans Affairs hospitals used race or ethnicity identification reported by employees or clinicians observing

patients; today they ask patients to report this information (Sohn et al. 2006). Analyses comparing racial and ethnic self-identification versus identification assigned by observers found excellent agreement but with some notable exceptions (among American Indians, Alaska Natives, and Hispanic whites). Instructions for birth registrations stipulate that personal information be obtained from mothers, fathers, or other knowledgeable persons (Hahn, Mulinare, and Teutsch 1992). In contrast, death certificates typically use statements from the next of kin to funeral directors, although funeral directors sometimes independently assign race according to their views of the decedent's appearance. Reliance on the opinion of funeral directors or sources unrelated to decedents yields highly suspect data on race. Studies have found substantial changes in death rates by racial group (e.g., among American Indians) when racial misclassifications are corrected (Williams 1999, 127).

Some have advocated abandoning collection of racial and ethnic data because of such ambiguities. Others counter that these data are crucial because social, economic, and political dynamics are inextricably linked to race and ethnicity by US history and persistent racism. Many physicians see no reason to gather these data on their patients, asserting that race and ethnicity are clinically irrelevant (Wynia, Ivey, and Hasnain-Wynia 2010). However, recognition of the physiological consequences of discrimination (e.g., effects on the hypothalamic-pituitary-adrenal corticoid axis, the so-called physiological stress response) and its related persistent stress on the health of individuals and populations is growing (Ren, Amick, and Williams 1999; Braveman, Egerter, and Williams 2011; Chae et al. 2010; Mujahid et al. 2011). Perceptions of racism could be related to higher likelihood of engaging in risky health behaviors, such as smoking, binge drinking, and unhealthy eating (Shariff-Marco, Klassen, and Bowie 2010). Persons who experience racism may engage in these risky behaviors to eliminate stress (Jackson, Knight, and Rafferty 2010). Improving the collection of systematic information on perceptions of race- or ethnicity-based discrimination might enhance our understanding of how discrimination affects individual and population health (Kressin, Raymond, and Manze 2008).

Motivated by the desire to gather and report information on health care disparities, Section 4302 of the ACA requires the collection of reliable information on a variety of subpopulations, including racial and ethnic minorities (and persons with disabilities; see Chapter 15). In June 2011, the Office of Minority Health (OMH) issued a proposed rule for collecting information on race and ethnicity and other designated population groups in federal surveys. In response to public comments, OMH released its final rule on October 31, 2011 (see Exhibit 3.5).

To address concerns about data accuracy, some have advocated imputation of race and ethnicity using other sources (Weissman and Hasnain-Wynia 2011, 2276–77):

The most common indirect method uses geographically coded data from the U.S. Census to characterize people on the basis of their address or ZIP Code as living in a high-, medium-, or low-minority area. A second approach uses each person's surname, along with Census information on the self-identification of people with that name. Surnames are especially useful for identifying Hispanics and Asians, whereas "geocoding" is most useful for identifying blacks.

Combining geocoding and surnames can substantially increase imputation's accuracy. For example, someone named Smith living in an area with a high proportion of blacks is more likely to be black than someone named Smith living in a largely white community. The newer indirect methods do not assign a single race or ethnic background to any individual; instead they estimate probabilities for each race or ethnic category.

I AND II. RACE AND ETHNICITY

Ethnicity Data Standard	Categories
Are you Hispanic, Latino/a, or Spanish origin? (One or more categories may be selected.) a. ____ No, not of Hispanic, Latino/a, or Spanish origin b. ____ Yes, Mexican, Mexican American, Chicano/a c. ____ Yes, Puerto Rican d. ____ Yes, Cuban e. ____ Yes, another Hispanic, Latino, or Spanish origin	These categories roll up to the Hispanic or Latino category of the OMB standard
Race Data Standard	**Categories**
What is your race? (One or more categories may be selected.) a. ____ White b. ____ Black or African American c. ____ American Indian or Alaska Native	These categories are part of the current OMB standard
d. ____ Asian Indian e. ____ Chinese f. ____ Filipino g. ____ Japanese h. ____ Korean i. ____ Vietnamese j. ____ Other Asian	These categories roll up to the Asian category of the OMB standard
k. ____ Native Hawaiian l. ____ Guamanian or Chamorro m. ____ Samoan n. ____ Other Pacific Islander	These categories roll up to the Native Hawaiian or Other Pacific Islander category of the OMB standard

Source: OMH (2011).

EXHIBIT 3.5 Data Collection Standards for Race and Ethnicity in Federal Surveys

Health services researchers often include race in their analyses as a proxy for the unmeasured quantities of socioeconomic status, discrimination, culture, and unspecified but presumed biological differences (LaVeist 1994). Williams (1994) examined 192 studies published between 1966 and 1990 in the journal *Health Services Research* and found that 63 percent considered race and ethnicity. However, 54.5 percent made black–white distinctions only (most included this binary variable as a dummy in regression equations), and only 13.2 percent defined or justified the use of race in their research. The use of racial variables in regression analyses may obscure important relationships, such as those involving education, socioeconomic class, culture, and health beliefs (Schulman et al. 1995). Other complicating factors include persons' primary language, English proficiency, country of origin, and immigration status (Williams 2005; Williams et al. 2010).

Consideration of race and ethnicity is further complicated by disparities in the quality of health care between racial and ethnic minorities and white populations, as extensively documented in *Unequal Treatment* (IOM 2002a). Review of the voluminous literature on health care use and outcomes across racial and ethnic groups is beyond the scope of this text; the *National Healthcare Disparities Report*, published annually by congressional mandate since 2003 by the Agency for Healthcare Research and Quality (AHRQ), describes the extent to which disparities in care exist not only for racial and ethnic minorities but also for other potentially vulnerable populations. The most recent edition (AHRQ 2011a) demonstrates that although some disparities are narrowing, others are worsening. For instance, in 2009, compared with whites, colon cancer screening rates worsened among blacks, Asians, American Indians and Alaska Natives, and Hispanics; similarly, disparities in the number of hospital patients with heart failure who received recommended hospital care grew among American Indians/Alaska Natives and Hispanics.

Concerns about differential quality of care raise serious questions about adjusting for race or ethnicity in examining outcomes of care: Such adjustment could hide significant racial disparities or confuse interpretation of the results. Some data collection systems, especially among private health insurers, do not collect information on race or ethnicity; their users argue that gathering such data perpetuates discriminatory attitudes. However, as Williams (1994) observed, evidence of persisting racial bias in the health care system demands continued scrutiny. Examination of outcome differences by race and ethnicity is critical to understanding fully the effects of this important patient attribute. The best strategy is to examine results separately for patients grouped into strata by race and ethnicity.

Acute Clinical Stability

Acute clinical stability reflects patients' current physiological functioning as indicated by basic homeostatic measures, such as vital signs (heart rate, respiratory rate, blood pressure, and temperature), serum electrolytes, hematologic findings (e.g., hematocrit, white blood cell count, clotting measures), arterial oxygenation, and levels of consciousness or neurologic functioning. The goal is to assess whether patients face an imminent risk of death. For example, although Patient G's colon cancer is more extensive than Patient F's (see Exhibit 3.2), Patient F faces immediate, life-threatening risks from sepsis.

Assessment of acute clinical stability is crucial to studying outcomes of acutely ill patients in short time frames, such as death during an acute care hospitalization or within a brief time window (e.g., 30 days from admission). The physiological risk factors represent basic organ functions required to keep patients alive (e.g., cardiac, respiratory, renal, neurologic function). Concentrating on core physiological functions makes this dimension generic or similar across specific diagnoses; at a certain point, a "final common pathway" of physiological dysfunction unfolds. This predictable progression derives from the concept of homeostasis expounded by Walter B. Cannon in 1929: "The body's major physiological systems interact to maintain internal balance and rapidly correct disturbances" (Wagner, Knaus, and Draper 1986, 1389). In other words, regardless of whether the patient develops shock because of sepsis, heart failure, or hemorrhage, in the final stages physiological indicators (e.g., heart rate, blood pressure) become deranged in similar ways.

Over more than three decades, various research groups and organizations have developed risk adjustment methods relating to patients treated in ICUs. These methods typically have indicators of acute physiological function at their core. The earliest of these ICU measures (see Exhibit 3.6) were major advances in the risk adjustment field and shared several features. First, a relatively small set of variables supplied the most important predictors across both children and adults. While the first version of APACHE contained information on 34 physiological parameters (Knaus et al. 1981), in producing APACHE II, Knaus and colleagues (1985) found that only 12 physiological parameters retained acceptable statistical performance. In producing APACHE III, Knaus and colleagues (1991) added several variables to the core group. The Simplified Acute Physiology Score (SAPS) II uses fewer variables (Le Gall, Lemeshow, and Saulnier 1993), and the Mortality Probability Model (MPM) is even more economical (Lemeshow et al. 1993, 1994). PRISM also uses few variables to predict outcomes in pediatric ICUs (Pollack, Ruttimann, and Getson 1988). The Sequential Organ Failure Assessment (SOFA), developed by the European Society of Intensive Care Medicine, contains only six variables (Vincent et al. 1996, 1998; Ferreira et al. 2001).[7]

EXHIBIT 3.6
Physiological Variables Included in Early Risk Adjustment Measures for ICU Patients

	APACHE			Mortality Probability Model II		
Clinical Variable	II[a]	III[b]	PRISM[c]	Admission[d]	At 24, 48, and 72 hours[e]	SAPS II[f]
A-a gradient	X	X				
Albumin		X				
Arterial carbon dioxide		X	X			
Arterial oxygenation	X	X	X		X	
Arterial pH	X	X				
Blood pressure	X	X	X	X		X
BUN		X				X
Bicarbonate			X			X
Calcium			X			
Glucose		X	X			
Heart rate	X	X	X	X		X
Hematocrit	X	X				
Level of consciousness, Glasgow Coma Score	X	X	X	X	X	X
PaO_2/FiO_2						X
Prothrombin time/ partial thromboplastin time			X		X	
Respiratory rate	X	X	X			
Serum creatinine	X	X			X	
Serum sodium	X	X				X
Serum potassium	X		X			X
Temperature	X	X				X
Total bilirubin		X	X			X
Urine output		X			X	X
Pupillary reactions			X			
White blood cell count	X	X				X

[a]Knaus et al. (1985).
[b]Knaus et al. (1991).
[c]Pollack, Ruttimann, and Getson (1988).
[d]Lemeshow et al. (1993).
[e]Lemeshow et al. (1994).
[f]Le Gall, Lemeshow, and Saulnier (1993).

Second, because these physiological parameters serve as clinical guideposts for physicians treating acutely ill patients, they are measured routinely and with minimal technological intervention. Since relatively few values are unmeasured, analysts avoid vexing questions about handling missing values. ICUs monitor routine physiological parameters using automated probes and other devices. Data downloaded directly from monitoring equipment can produce credible acute physiology scores (Junger et al. 2002).

Third, measures of acute clinical stability are used most commonly to predict risk of imminent death for gravely ill patients. Acute physiological status, however, may not sufficiently predict imminent death. MPM also considers the presence of metastatic cancer, chronic renal failure, and cirrhosis (Lemeshow et al. 1993); SAPS II incorporates age and metastatic cancer, hematologic malignancy (lymphoma, acute leukemia, multiple myeloma), and acquired immunodeficiency syndrome (AIDS) (Le Gall, Lemeshow, and Saulnier 1993). APACHE III adds an age category and points for seven comorbid conditions (Knaus et al. 1991).

Finally, recent acute physiology–based risk adjustment methods have been developed empirically with clinical conceptual guidance (see Chapter 8). This approach raises various questions. Empirically derived methods inevitably reflect treatment patterns and outcomes from their underlying databases. Intensive care interventions are constantly changing, however, and new therapeutic techniques sometimes change patients' outcomes dramatically. In the last decade, for instance, wider use of low tidal volume ventilation and the introduction of drotrecogin alfa (activated) for treating sepsis were among the innovations that led to the development of APACHE IV. Given these new ICU practices, the older version no longer predicted outcomes accurately (Zimmerman et al. 2006). Changes in palliative care practices also motivated the update; APACHE IV contains considerably more information about chronic conditions and diagnoses than did the earlier versions (Zimmerman and Kramer 2008). With a data set containing more than 130,000 ICU admissions from 45 hospitals, the developers took advantage of powerful new statistical modeling approaches, expanding the use of splines for certain component variables.

The developers of APACHE IV view its excellent predictive performance and its complexity as a departure from the early ICU risk adjusters that used relatively few predictive variables and transparent algorithms (i.e., algorithms open to user review) to compute scores (Zimmerman et al. 2006, 1305):

> The increased complexity of APACHE IV represents a continued departure from the simplification that characterized APACHE II and other second-generation prognostic scoring systems. A recently developed automated risk adjustment system for Veterans Affairs ICUs is similar to APACHE IV in its complexity and emphasis on information technology. We also know that the accuracy of older prognostic models such as APACHE II has deteriorated over time, and they lack predictor variables of proven prognostic significance.

The new generation of ICU risk adjusters is being used to measure and benchmark ICU performance, concerning not only mortality but also length of stay. A growing number of studies are comparing the predictive performance of ICU risk adjusters within different clinical contexts and for different outcomes (Afessa, Gajic, and Keegan 2007; Bakhshi-Raiez et al. 2007; Higgins 2007; Higgins, Teres, and Nathanson 2008; Minne, Abu-Hanna, and de Jonge 2008; Vasilevskis et al. 2009; Shrope-Mok, Propst, and Ivengar 2010). The relative performance of different ICU risk adjusters depends on the setting and outcome of interest.

The value of acute physiological parameters in assessing patient risk for imminent clinical outcomes is undisputed. Acute clinical stability is central to evaluating risk for imminent death and complications, such as respiratory failure and shock. Some methods are specifically designed to predict length of stay. As ICU therapies continue to evolve, so too must specific risk adjusters that model acute physiology. The availability of electronic health information from patient monitoring in ICUs and new laboratory tests will also drive these changes.

Diagnoses and Health Conditions

> Diagnosis is the focal point of thought in the treatment of a patient. From diagnosis, which gives a name to the patient's ailment, the thinking goes chronologically backward to decide about pathogenesis and etiology of the ailment. From diagnosis also, the thinking goes chronologically forward to predict prognosis and to choose therapy. As the main language of clinical communication, diagnostic labels transmit a rapid understanding of the contents of the package . . . (Feinstein 1967, 73)

From resource use to clinical events, risks for diverse outcomes differ dramatically by diagnosis. Definition of diagnoses has preoccupied medicine since Hippocratic times, spawning volumes of nomenclatures with specifications delineating disease characteristics. Typically, diagnosis requires specification of the organ or organ system affected, indications of the pathophysiological abnormalities wrought by the disease process, and identification of the etiology or cause (or set of causes) producing the pathophysiological changes. The term *health condition* is often used when the dimensions that define diagnoses are not completely specified.

Sometimes, delineating diagnoses along these dimensions is straightforward (e.g., pneumonia infection of the lung caused by *Streptococcus pneumoniae*). Other times, definitive diagnoses are impossible to make. For example, despite fiber-optic studies, angiography, radionuclide scans, and barium enemas, the source and causes of lower gastrointestinal tract bleeding may remain elusive. Clinicians assign a "diagnosis" of gastrointestinal hemorrhage, reflecting a sign of underlying disease rather than the disease itself. Furthermore, the demands of acutely managing a potentially cataclysmic

event may divert attention from precise diagnosis. Later, after stabilizing the patient, clinicians can investigate the exact cause.

In certain situations, the dividing line between the presence and absence of disease is blurred. For example, dysplasia of the cervix, an abnormal finding on a Papanicolaou test, can presage cervical cancer. Subtle morphologic differences between dysplastic and normal cells compromise the reliability of the test, and the condition generally vanishes spontaneously. Even while present, dysplastic cells do not cause pain or disability. Numerous examples exist of "diseases . . . being defined by an abnormal result on some test, leaving uncertainty about its real meaning to a patient and the appropriate treatment" (Eddy 1984, 76).

Vague diagnostic terminology may reflect habit, sloppiness, or situations in which specifying the cause is clinically unimportant. For example, heart failure affected about 5.8 million persons in the United States and cost the country approximately $39.2 billion in 2010 (CDC 2010a). The most common causes are long-standing hypertension, coronary artery disease, and diabetes, but heart failure sometimes results from such diverse etiologies as viruses, alcohol, or disordered iron metabolism. Some causes require specific interventions. Once heart failure is present, however, acute exacerbations are generally treated similarly. Therefore, physicians may not specify the cause of heart failure when listing this diagnosis.

Finally, the manifestations of a single diagnosis may vary widely. For example, complications of diabetes range from blinding retinopathy to chronic renal failure, each requiring different diagnostic evaluations and therapeutic interventions. Similarly, diverse diseases may lead to similar problems. For example, end-stage renal disease may result from hypertension, diabetes, lupus, or certain infections; end-stage renal failure from any cause requires renal transplantation or lifelong dialysis. Depending on the purpose of the analysis, it may be appropriate to group all patients with certain complications regardless of underlying etiology.

When using diagnosis as a risk factor, having a definitive diagnosis that meets rigorous standards may not be essential, depending on the context. Such vagaries often reflect the realities of today's clinical practice. However, questionable diagnostic information (e.g., diagnoses that are not definitively established) may affect the utility of diagnosis as a risk factor.

Severity or Extent of Diagnoses

In some contexts, knowing principal diagnoses alone is insufficient; analysts must also understand the severity or extent of the diagnoses. For example, persons with a nidus of well-differentiated colon cancer isolated to a polyp (patients C to F, Exhibit 3.2) have different prognoses and therapeutic needs than persons with widely disseminated malignancy (patients J and K). Many studies of medical effectiveness and patient outcomes focus on a single principal diagnosis, making

it unnecessary to separate patients by diagnoses. Nevertheless, determination of severity levels within the diagnostic category often remains important.

The concept of severity has many layers, commonly organized around prognoses or expectations about patients' clinical outcomes driven by the extent and nature of diseases. Prognoses vary depending on the time frame—months or years versus hours, days, or weeks (see Chapter 4)—thus long- and short-run assessments of severity may differ. For example, Patient F (Exhibit 3.2) is septic as a result of bloodstream infection by dangerous bacteria. The patient needs immediate, aggressive treatment. Later, a colonoscopy reveals an adenomatous polyp containing a small nidus of adenocarcinoma. In contrast, Patient K's colon cancer has metastasized widely. The patient needs treatment to alleviate pain and appears cachectic but is not acutely ill. The short- and long-term prognoses of these patients differ. The first has an immediate life-threatening condition; if treated successfully, the patient could live for years. The second patient's condition is not immediately life threatening but portends a poor long-term prognosis.

The term *prognosis* generates the same question as does the word *risk*: Prognosis for what outcome over what time frame? Many studies and inpatient hospital performance measures use death as their outcome of interest, but it is most relevant for diagnoses that commonly result in imminent death regardless of intensive therapeutic intervention. For assessing most outpatient conditions, however, imminent death is irrelevant. For instance, a study of nearly 350,000 men tracked for up to 25 years found that individuals with stage 2 hypertension (systolic values greater than or equal to 160 mmHg or diastolic values greater than or equal to 100 mmHg) had adjusted hazard ratios of all-cause mortality of 1.44 (95 percent confidence interval 1.28, 1.62) (Terry et al. 2007). Therefore, are men with stage 2 blood pressure levels 44 percent more "sick" than those with normal pressures (systolic values lesser than 120 mmHg and diastolic values lesser than 80 mmHg)? Because death is often years away, this diagnosis may make little difference in the immediate assessment of the patient, depending on the overall clinical context.

Diseases with protean presentations introduce special complexities. Which is the most severe manifestation of diabetes: blindness, end-stage renal disease, or debilitating peripheral vascular disease? Quality-of-life considerations enter this equation, albeit tied to personal values (e.g., one person may adapt to blindness, whereas another may find it intolerable). Acute versus chronic manifestations of single diseases make comparison of severity even muddier. For example, is a patient whose sickle-cell anemia produces periodic painful crises sicker than one whose disease caused renal failure? Comparison of severity among diseases is yet more challenging, often amounting to "comparing apples with oranges." At the outset, colon cancer seems more serious than psoriasis, for example, but severity depends on the extent of each

disease. Early, asymptomatic colon cancer is less severe (e.g., in terms of imminent death, discomfort, disability) than psoriasis with diffuse erythroderma and an antibiotic-resistant bacterial infection.

Complete measurement may require observing patients over time. For example, suppose three patients experience identical acute neurological deficits as a result of cerebrovascular disease. Patient 1's symptoms resolve fully in 24 hours; Patient 2's deficits slowly improve over a week but do not vanish completely; and Patient 3's debilities persist without change. The severity of cerebrovascular disease deficits clearly differs among the three patients, but observers focusing only on their initial impairments miss these distinctions. However, as discussed in Chapter 4, one pitfall of longitudinal examinations is the potential to confound outcomes with substandard care.

Finally, diagnosis may affect how other risk factors affect outcomes. For example, a temperature of 102°F might be very severe for patients with leukemia but only moderately severe for immunocompetent patients with pneumonia (Horn et al. 1991; Iezzoni and Daley 1992). As noted earlier, recent generations of measures of acute physiological severity are incorporating more information about patients' diagnoses.

Comorbidities

Especially as persons age, the likelihood of having more than one diagnosis or health condition increases. In 2005, among US noninstitutionalized residents aged 0 to 19 years, 16.5 percent had one, 3.7 percent had two, and 1.2 percent had three or more chronic conditions; among those aged 65 to 79 years, 20.2 percent had one, 21.5 percent had two, and 45.3 percent had three or more chronic conditions (Paez, Zhao, and Hwang 2009). The increasing prevalence of coexisting chronic health problems has generated a new word (*multimorbidity*) and its own acronym: MCCs (multiple chronic conditions).

Comorbidities, or coexisting diagnoses, are diseases with unrelated etiologies (i.e., causalities). Comorbidities differ from *complications*, which are sequelae or predictable consequences of a particular diagnosis. For example, for colon cancer patients (Exhibit 3.2), cerebrovascular disease (Patient D) is a comorbidity, whereas bowel obstruction (Patient I) is a complication. The prototypical comorbidity is a chronic condition, such as diabetes mellitus, chronic obstructive pulmonary disease, or chronic ischemic heart disease. Depending on the context, however, comorbid illnesses may be acute (e.g., heart attack during admission to treat colon cancer). Complications and comorbidities are sometimes difficult to distinguish. For example, Patient E's serious depression is a comorbidity of his early-stage colon cancer. However, persons diagnosed with cancer might experience depression as a result of this life-transforming event; in this case, depression would be considered a complication.

Acute comorbidities, such as heart attacks or pneumonia, may be iatrogenic. Questions about quality of care and medical errors arise when these conditions are listed as secondary diagnoses during hospital stays. The semantic distinctions among acute comorbidities, complications, iatrogenic illnesses, adverse events, and other similar terms are often unclear. Although conceptually important, distinguishing acute comorbidities related to underlying disease from those caused by iatrogenic events is difficult in many cases. One practical problem is determining the timing of events; some data sources offer little or unreliable insight into when events occurred. Addition of the present on admission (POA) flag to discharge diagnoses is intended to assist in distinguishing newly occurring conditions from those existing prior to hospitalization (see Chapter 5). As of October 1, 2008, Medicare stopped paying hospitals for treating the following conditions if patients acquire any of them after admission: air embolism; blood incompatibility; catheter-associated urinary tract infection; pressure ulcer; object left in patient during surgery; vascular catheter-associated infection; mediastinitis after coronary artery bypass grafting; and fall from bed (Rosenthal 2007). This policy change assumes that these acute conditions are complications of care.

In most instances, compared to persons without MCCs, patients with multiple comorbidities have higher risks of death and complications and have higher rates of functional impairments and disability. They may generate higher health care costs because they require additional diagnostic testing and therapeutic interventions. The effect of comorbidities can vary by time frame; MCCs strongly influence long-term survival and functioning. Most randomized controlled trials exclude patients with MCCs because of concerns that comorbid disease might confound assessments of treatment efficacy. However, observational and medical effectiveness studies examine outcomes in the real world, where chronic conditions are common, especially as patients age. Therefore, comorbidities are potentially critical risk factors, especially when examining adult populations.

Despite this potential importance, inclusion of comorbidities in risk adjustment has presented considerable methodological challenges. Approaches include (1) assigning a single score representing combined effects from all health conditions, (2) adding points for individual health conditions to produce a single score, and (3) entering variables for each condition into multivariable regression models and allowing each to generate its own weight within the specific data set (see Chapter 10). An example of the first approach is the Physical Status Classification of the American Society of Anesthesiologists (ASA), used in preoperative evaluations of surgical patients (Cohen et al. 2009; Dimick et al. 2010; Hightower et al. 2010). ASA scores rate risks of perioperative death on a global, subjective, five-point scale encompassing all aspects of a patient's presentation (level 5 indicates patients expected to die within 24

hours). Some have raised concerns that the ASA scores represent subjective judgments and could be inflated to make surgical patients appear sicker than they actually are—a concern when this information is used to assess hospital performance (Cohen et al. 2009).[8]

The Charlson Comorbidity Index exemplifies the second method: adding points for individual health conditions. More than two decades ago, Charlson and colleagues (1987) developed this index by initially examining one-year survival among 559 patients admitted to the medical service at New York Hospital during a one-month period in 1984. For each of 19 candidate conditions, they calculated adjusted relative risks from a Cox proportional-hazards model and then assigned weights ranging from 1 to 6. The weighted index significantly predicted one-year mortality ($p < 0.0001$). Investigators later used coded data from hospital discharge abstracts to adapt the Charlson model for applications that use such data (Deyo, Cherkin, and Ciol 1992; Romano, Roos, and Jollis 1993). Exhibit 3.7 compares the variables included in the Charlson method and their assigned weights with those included in the Elixhauser method, an increasingly prominent comorbidity measure described later in the chapter.

Another comorbidity measure that added points for individual conditions was developed by RAND investigators for their study of the effect of Medicare's prospective hospital payment system (Keeler et al. 1990). In addition to considering diseases, the RAND researchers employed broad definitions of comorbidities, including smoking, obesity, and certain procedures (e.g., use of a nasogastric tube, thoracic or abdominal surgery in the last month, use of home oxygen). The researchers assigned weights to each of 16 conditions using clinical judgment and logistic regression predictions of 30-day mortality. Their comorbidity index significantly predicted 30-day post-admission mortality for only two of the five study conditions (pneumonia and hip fracture) when added to a model containing acute physiological variables.

Given the widespread availability of large databases, use of the third type of comorbidity measures—those that assign empirical weights through multivariable modeling—is becoming increasingly prevalent. Two major examples are methods developed by Elixhauser and colleagues (1998) from AHRQ (Exhibit 3.7) and the Hierarchical Condition Categories (HCCs) (Pope et al. 2004). Using coded hospital discharge abstract data, Elixhauser and collaborators identified 30 conditions associated with length of stay, hospital charges, and in-hospital deaths. Developers of the HCCs sorted inpatient and outpatient diagnosis codes into 189 condition categories, creating hierarchies of conditions by severity within broad diagnostic groups. In both instances, investigators enter these variables into their own models, allowing each condition variable to generate its own weight during multivariable regression. Depending on the purpose and the data set, different

EXHIBIT 3.7 Variables Included in Charlson and Elixhauser Comorbidity Measures

Charlson Comorbidity Indicators	Elixhauser Comorbidity Indicators
Congestive cardiac insufficiency (weight 1)	Congestive heart failure
Myocardial infarction (weight 1)	
	Valvular disease
	Pulmonary circulation disorders
Peripheral vascular disease (weight 1)	Peripheral vascular disease
	Hypertension (combine uncomplicated and complicated)
Hemiplegia (weight 2)	Paralysis
Cerebrovascular disease (weight 1)	
	Other neurological disorders
Chronic pulmonary disease (weight 1)	Chronic pulmonary disease
Diabetes without complications (weight 1)	Diabetes without chronic complications
Diabetes with complications (weight 2)	Diabetes with chronic complications
	Hypothyroidism
Moderate or severe kidney disease (weight 2)	Renal failure
Chronic diseases of the liver or cirrhosis (weight 1)	Liver disease
Moderate or severe liver disease (weight 3)	
Ulcers (weight 1)	Chronic peptic ulcer disease (includes bleeding only if obstruction is also present)
AIDS (weight 6)	HIV and AIDS
Leukemia (weight 2)	
Lymphoma (weight 2)	Lymphoma
Malignant tumors, metastases (weight 6)	Metastatic cancer
Tumors (weight 2)	Solid tumor without metastases
Connective tissue disease (weight 1)	Rheumatoid arthritis/collagen vascular diseases
	Coagulation deficiency
	Obesity
	Weight loss
	Fluid and electrolyte disorders
	Blood loss anemia
	Deficiency anemias
	Alcohol abuse
	Drug abuse
Dementia (weight 1)	
	Psychoses
	Depression

Source: Adapted from HCUP (2011).

methods (or combinations of methods) perform better than others in predicting the outcome of interest (Southern, Quan, and Ghali 2004; Li et al. 2008; Maciejewski, Liu, and Fihn 2009; Fleishman and Cohen 2010; Li, Kim, and Doshi 2010; Gagne et al. 2011).

Functional Status

In the United States at least 54 million persons live with disabilities or some type of functional impairment (AHRQ 2010b). This figure will increase substantially in coming decades as baby boomers age (IOM Committee on Disability in America 2007). One longitudinal study found that only 17 percent of persons experienced no disability in their last year of life (Gill et al. 2010). An increasing number of youth are also living with functional impairments, many assisted by technological breakthroughs and major advances in medical science. Worrisome trends linked primarily to obesity among children and young adults (Alley and Chang 2007) suggest that the number of middle-aged persons with impaired functioning might also increase.

According to 2001–2005 National Health Interview Survey data, 29.5 percent of US civilian, noninstitutionalized adults experience difficulties with basic actions: 21.7 percent have movement difficulties (problems with walking, standing, kneeling or bending, reaching overhead, and using hands or fingers); 3.1 percent have emotional difficulties; 13.1 percent have difficulties seeing or hearing; and 2.8 percent have cognitive difficulties (Altman and Bernstein 2008). Many persons report more than one type of difficulty. Among persons aged 18 to 44, 6.8 percent report complex activity limitations (difficulties with social activities, work, and self-care) and 16.7 percent report difficulties with basic actions (problems with movement or sensory, emotional, or cognitive functions). Among persons aged 65 or older, 36.3 percent report complex activity limitations and 33.1 percent report difficulties with basic actions. Women report higher rates of problems than do men. Functional status problems may be higher among certain racial and ethnic groups; American Indian or Alaska Native adults report the highest disability rates (30 percent), followed by blacks (21 percent), whites (20 percent), Hispanic individuals and Native Hawaiian and other Pacific Islanders (17 percent), and Asians (12 percent) (CDC 2008). Given that these numbers represent self-reports it is unclear whether cultural factors relating to willingness to reveal impairments might bias these figures.

According to the World Health Organization (2001) in its *International Classification of Functioning, Disability and Health* (ICF), the concept of disability melds "biopsychosocial" perspectives. ICF draws on three interrelated concepts: "Impairments are problems in body function or structure such as a significant deviation or loss. Activity is the execution of a task or action by an individual. Participation is involvement in a life situation." ICF defines disability

as an "umbrella term for impairments, activity limitations or participation restrictions," conceiving "a person's functioning and disability . . . as a dynamic interaction between health conditions (diseases, disorders, injuries, traumas, etc.) and contextual factors," including environmental and personal attributes. The introduction of environmental factors, such as physical barriers, as determinants of disability is a novel aspect of ICF's model.

In comprehensive risk assessments, functional status is distinguishable from the concepts of health status and quality of life described later in this chapter. Some find such distinctions spurious, however, preferring global measures of well-being that cut across various dimensions. Nonetheless, an important distinction is that functional status typically captures observable behaviors rather than individuals' perceptions of health (Rubenstein et al. 1989, 563):

> Functional status . . . encompasses the more limited areas of physical, mental, and social functioning in daily life. Functioning is observable; it consists of everyday behaviors as they occur in a person's home and community life. Measures of functioning include items about daily activities such as eating, dressing, bathing, walking, handling finances, or visiting friends and relatives. Functional status is the end result of a person's health (absence of disease), well-being (capacity to participate fully in life), and coping (capacity to overcome health problems).

Hundreds of measures quantify functional status within specific clinical conditions or generically across diagnoses. New computerized techniques, such as those using item response theory (Chapter 7), enable surveyors to elicit specific information by asking small numbers of well-targeted questions. Historically, functional status measures have typically included basic activities of daily living (ADLs; e.g., feeding, bathing, dressing, toileting, walking) and instrumental ADLs (IADLs; e.g., shopping, preparing meals, doing housework, using public transportation, using the telephone, balancing a checkbook). More comprehensive measures of functioning also address cognitive abilities (e.g., level of alertness, orientation, long- and short-term memory, capacity for learning and computation), affective health (e.g., happiness, anxiety, depression), and social activities (e.g., visiting friends, sexual relationships).

Launched in 2002, the Patient-Reported Outcomes Measurement Information System (PROMIS) is a network funded by NIH that is compiling sets of dynamic tools aimed at capturing patient-reported functional status and other outcomes. According to its website (www.nihpromis.org), the goal of this initiative is to build and validate "common, accessible item banks to measure key symptoms and health concepts applicable to a range of chronic conditions, enabling efficient and interpretable clinical trial research and clinical practice application of patient-reported outcomes." These item banks are undergoing intensive psychometric testing, including testing by an Internet polling panel (Liu et al. 2010). To organize its extensive item banks, PROMIS groups its measures into domains and uses item response theory to

analyze them (Cella et al. 2007; Hays et al. 2007). According to the August 2010 draft of these domains (posted on the website), physical health items are sorted into two domains: (1) symptoms and (2) function, which includes physical function, sexual function, sleep function, and physical activity. The social health domain also includes a branch for function, which sorts measures into two groups (ability to participate and satisfaction with participation). Initial studies suggest that the PROMIS item banks produce reliable ratings of function that are comparable to the ratings produced by legacy instruments (Cella et al. 2010).

Although item banks like PROMIS suggest that assessing functional status is straightforward, complexities abound. First, the measurement context can affect perceptions of functional ability. According to Young and colleagues (1996), "*capability* indicates what persons can do in controlled settings, whereas *performance* assesses what persons do in everyday life. Capability typically exceeds performance." Second, patients' perceptions of their functional ability often differ from their clinicians' perceptions. In the Framingham Heart Study initiated in 1948, about 1,400 participants completed a disability questionnaire before their biennial physical examination. During the examination, nurses tested these individuals' functional abilities (Kelly-Hayes et al. 1992). Most of the roughly 7 percent of differences found between self-reported and observed performance pertained to walking and stair climbing; for at least 89 percent of discrepancies, respondents reported significantly worse functioning than study nurses observed.

In contrast, a study of 620 women who had recent strokes found that the women reported considerably better functioning than was found on physical testing; self-reports and examination results disagreed substantially for 19.3 percent of women and slightly for 55 percent (Owens et al. 2002, 806). The researchers suggested that physicians should conduct physical performance tests rather than rely on patients' self-reports. Many physicians, however, are poor at assessing patients' functional limitations. A study of 408 outpatients found that 22 percent reported difficulties walking one block or climbing one flight of stairs; 31 percent had trouble walking several blocks. Their 118 physicians underestimated or failed to recognize roughly two-thirds of these problems (Calkins et al. 1991).

Another question is whether the mode of collecting functional status information (e.g., face-to-face interview, self-administered mail survey, telephone interview, Internet) introduces bias. In face-to-face interviews, respondents may hesitate to reveal the extent of their dysfunction. A study that administered the 36-item Short Form Health Survey (SF-36) to 172 veterans found significant differences in patients' reports over the course of a week, depending on the mode of administration. For four of the eight SF-36 dimensions, face-to-face administration elicited a more optimistic view of health than did self-administration (Weinberger et al. 1996). More research

is needed to evaluate whether Internet administration produces results comparable to those produced by other modes of administration.

Depending on the population of interest and the functional status measurement tool, concerns about floor and ceiling effects also arise. Bindman, Keane, and Lurie (1990) used the Medical Outcomes Study (MOS) general health survey (MOS-20), an instrument largely designed and tested among outpatients, on 414 patients in poor health admitted to public hospitals. Six months later, the patients were asked if their health had changed. At baseline, the patients had much lower functional status than that previously reported for outpatient and general population cohorts; according to the survey results, these poor functional levels changed little over six months. Nonetheless, more than half of the public hospital patients reported that their health status had declined. In-depth evaluation suggested that the MOS-20 failed to detect declining health among very sick patients likely as a result of the *floor effect* (i.e., because these patients had initially rated themselves at the lowest level allowed by the MOS-20, their subsequent ratings could not fall). The opposite concern, a *ceiling effect*, arises when instruments fail to detect improvements in the outcome of interest (e.g., functional abilities) among persons with high levels of that attribute.

Baseline functional status strongly predicts subsequent functional status. Functional status also significantly predicts other outcomes, such as imminent death, complications, resource consumption, and satisfaction with care. Among Medicare beneficiaries with pneumonia, baseline walking difficulties significantly predicted mortality 30 days following hospitalization (Daley et al. 1988; Keeler et al. 1990). Functional status may predict health care costs, although it may be less predictive than information about diagnoses or chronic health conditions (Maciejewski, Liu, and Fihn 2009; Fleishman and Cohen 2010). As such, functional status measures supplement diagnostic and demographic information in determining capitation payment levels; one example is Medicare's use of a frailty adjuster to set capitated payment for its Program of All-Inclusive Care for the Elderly (Robinson and Karon 2000; Kautter, Ingber, and Pope 2008). Medicare beneficiaries with functional impairments report significantly higher rates of dissatisfaction with their health care experiences along a range of dimensions (Iezzoni et al. 2002; Jha et al. 2002).

Head-to-head comparisons of functional status measures with physiological findings have yielded interesting results. Using forward stepwise logistic regression to predict in-hospital death for pneumonia and stroke patients, Davis and colleagues (1995) found that nurses' assessments of patients' functional status were stronger predictors than most laboratory test values and comorbid diseases.[9] A global assessment by nurses of patients' needs for ADL assistance better predicted in-hospital death among AIDS patients than did three validated AIDS mortality measures (Justice et al. 1996). These findings are not surprising given that functional status reflects how persons are doing

overall by considering their full range of diagnoses and physiological status. Increasingly, researchers are treating functional status as both a critical predictor and outcome measure, thereby raising additional complexities (High et al. 2005; Jackson et al. 2006).

Psychological, Cognitive, and Emotional Functioning

Psychological, cognitive, and emotional functioning encompass such attributes as individuals' ability to appreciate and interact with their surroundings and other people, their capacity to understand information about their health and health care needs and to act productively on this information (self-efficacy), and their emotional ability to respond to their health conditions in ways that maximize their well-being. These factors particularly affect outcomes outside controlled institutional environments. In hospitals, clinical staff oversee all patient needs; for example, they give medications on schedule in proper dosages. In the community, however, psychological and cognitive problems can compromise patients' activity levels, self-care motivation, and perceptions and negatively affect outcomes. Patient E (Exhibit 3.2), who refused treatment for serious depression, is at greater risk of poor outcomes than is Patient C, who has identical colon cancer. Chapter 14 explores risk adjustment specifically for mental health conditions.

Many scales have attempted to capture mental health and cognitive dimensions of functioning. NIH's PROMIS banks (described earlier) contain numerous items relating to these attributes. For instance, PROMIS has separate data banks for items relating to anger, anxiety, and depression. Its cognition domain has subgroups representing cognitive abilities and self-efficacy. Hays and colleagues (2009) constructed a global mental health measure and evaluated it using responses from an Internet panel participating in PROMIS instrument testing. This global measure correlated strongly with other methods used to assess depression.

Some of the concerns raised about measuring functional status also apply to measurement of mental health. Scales of cognitive functioning can suffer floor and ceiling effects (i.e., inability to detect small changes in cognitive functioning at either end of the scale). In addition, patients' level of education affects most cognitive functioning scales. Older people in particular may feel threatened when asked to complete cognitive evaluations, such as the Mini-Mental State Examination, and may try to memorize basic recall words to maximize their scores (Kutner et al. 1992). Furthermore, according to Applegate, Blass, and Williams (1990, 1210), "[b]ecause many elderly persons with mild-to-moderate cognitive impairment often maintain their social skills in terms of superficial interactions, clinically important impairment may remain undetected."

The recent inclusion of questions about sexual orientation in selected federal and state health surveys has enabled population-based exploration of sexual minority health (Conron, Mimiaga, and Landers 2010; Dilley et al. 2010).[10] In 2003, Washington State began including a sexual orientation question in its annual Behavioral Risk Factor Surveillance System, a telephone survey administered by the states but coordinated by the national Centers for Disease Control and Prevention (Dilley et al. 2010). About 1 percent of men and 1 percent of women answered "don't know" to this question, and 1.2 percent of men and 1.6 percent of women refused to answer. Among men, 1.9 percent reported being gay and 0.9 percent bisexual; among women, 1.4 percent reported being lesbian and 1.6 percent bisexual. The adjusted odds ratios of having poor self-reported physical and mental health and activities limited by their health were higher for sexual minorities than for others. More detailed analyses of these data, focused on women, found that lesbian and bisexual women reported somewhat different patterns of self-reported mental distress and health problems, suggesting that these groups form distinct populations (Fredriksen-Goldsen et al. 2010). *Healthy People 2020* is the first incarnation of the decennial report that proposes national health priorities for the decade ahead to include a chapter on lesbian, gay, bisexual, and transgender health, recognizing the potential for these populations to experience health and health care disparities caused by historical stigmatization and discrimination (www.healthypeople.gov/2020).

Health-Related Behaviors and Activities

This category of potential risk factors includes such attributes as tobacco use, obesity and overweight, diet and nutrition, leisure time physical activity, sleep, excessive alcohol use, illicit drug use, seat belt use, and unsafe sexual practices. These behaviors/conditions may or may not be under a person's control. Genetics predisposes some persons to overweight and others to alcohol abuse. Aspects of the communities in which people live also contribute. For example, healthy eating is more difficult in neighborhoods that do not have adequate grocery outlets, and it is harder to engage in physical activity in communities without playgrounds, parks, and safe streets or sidewalks. Therefore, as stressed by *Healthy People 2020*, social and physical factors are determinants of health. Evidence about the health effects of many of these factors is overwhelming, especially their impact on population health. Coverage of the full breadth of evidence is beyond the scope of this text; the following paragraphs offer selected highlights on tobacco use, obesity and overweight, and alcohol consumption.

Tobacco use is associated with nearly one of every five deaths in the United States, or approximately 443,000 deaths per year (CDC 2011a). Smoking causes 90 percent of lung cancer deaths among men (80 percent among women), other cancers (such as cancers of the esophagus, kidney, stomach, and other organs), coronary artery disease, and chronic obstructive pulmonary diseases. Rates of smoking vary across populations. For instance, data from the National Survey on Drug Use and Health showed that in 2010 cigarette use among persons aged 12 or older was 35.8 percent for American Indians and Alaska Natives, 29.5 percent for whites, 27.3 percent for blacks, 21.0 percent for Hispanics, and 12.5 percent for Asians (SAMHSA 2011).

In national discourse, obesity and overweight have overtaken tobacco as the leading public health concern. This shift is not surprising; according to Stewart, Cutler, and Rosen (2009, 2252), "If past obesity trends continue unchecked, the negative effects on the health of the U.S. population will increasingly outweigh the positive effects gained from declining smoking rates." In 2009, rates of obesity, which is defined as body mass index (weight [kg]/height [m^2]) greater than or equal to 30, varied from 19 percent in Colorado to 34 percent in Mississippi (CDC 2010b). While the national rate was 27 percent, the rate for non-Hispanic blacks was 37 percent. Adults with less than a high school education were also disproportionately obese (33 percent). Among male adults, individuals with higher incomes have slightly higher obesity prevalence, while among female adults, obesity prevalence is significantly lower at higher income levels (Ogden et al. 2010b).

Childhood obesity and overweight have reached alarming levels. Approximately 17 percent of children and teens (12.5 million persons) are obese (CDC 2011b). In the 1960s, only about 5 percent of this age group was obese. As among adults, obesity is nonrandomly distributed among children and teens across population subgroups. For instance, most obese children and adolescents are not low income (Ogden et al. 2010a).

Obesity, but not overweight, is associated with excess deaths (Flegal et al. 2005), especially from cerebrovascular disease and cancers related to obesity (Flegal et al. 2007). Rates of disability related to overweight and obesity are increasing over time, particularly among racial and ethnic minorities (Seeman et al. 2010) and younger individuals (Alley and Chang 2007). In addition to promoting these deleterious health effects, obesity compromises quality of life and is societally stigmatized. This stigmatization is leading to increasing discrimination against obese individuals, such as in the workplace (Puhl and Heuer 2010).

The extent to which individuals can control their weight continues to be controversial. New research is finding, for example, that insufficient sleep affects the body's neuroendocrine response to lower food intake, reducing

the effectiveness of reduced food intake and caloric restrictions for losing weight (Nedeltcheva et al. 2010). Thus, persons who are unable to sleep sufficient hours may find weight loss more challenging, although more research is needed to confirm these effects. Another growing body of research is examining the implications that the unavailability of neighborhood retail food outlets offering healthy food has for weight levels in the local population (Rose et al. 2010).

Individuals should theoretically be better able to control their physical activity levels. Evidence suggests that, to achieve health benefits, adults should perform at least 150 minutes of moderate intensity or 75 minutes of vigorous intensity physical activity per week or an equivalent combination of the two (CDC 2010c). After controlling for various confounding factors, a study using a national cohort of men and women aged 60 or older found that self-reported leisure time physical activity was associated with significantly lower mortality risks (adjusted hazards ratio = 0.52 [0.35–0.78]) (Gillum and Obisesan 2010). Leisure time physical activity also may offer other benefits, such as reduced depression (Teychenne, Ball, and Salmon 2010). Despite these positive effects, only about 25 percent of adult Americans achieved the CDC target for leisure time physical activity in 2007–2008 (CDC 2010c). As mentioned earlier, nonrandomly distributed environmental factors, such as the presence of parks, playgrounds, safe walking routes, and recreation facilities, might affect individuals' ability to achieve physical activity goals.

While moderate alcohol consumption may offer health benefits (Mukamal 2010), excess alcohol consumption contributes to approximately 79,000 deaths per year in the United States, making it the third leading lifestyle-related cause of death (CDC 2011c). According to national surveys, more than half of the US adult population drank alcohol within the last 30 days, approximately 5 percent drank heavily, and 15 percent engaged in binge drinking. Alcohol consumption patterns are not random but are influenced by cultural and social factors. Individuals' patterns of alcohol consumption are closely related to those of their relatives and friends, the closest members of their social networks (Rosenquist et al. 2010). Alcohol use also is closely linked to tobacco use (Mukamal 2010).

This brief overview of health-related behaviors as risk factors neglects concerns that might be important, depending on the outcome of interest. The body of knowledge about effects of health-related behaviors and activities will continue to grow, especially with the focus of *Healthy People 2020* on social and environmental determinants of health. The major challenge to using this information in risk adjustment is availability of data. With some exceptions, such as tobacco use and body mass index, information on many of these behaviors is not routinely reported in administrative data systems or even medical records.

Social, Socioeconomic, and Environmental Factors

This broad category of risk factors considers social, socioeconomic, and environmental determinants of health at the levels of individuals and populations. Potential risk factors include familial characteristics, marital or partner status, and household composition; social networks; educational attainment and health literacy; employment and occupation; income and economic resources; housing and homelessness; neighborhood characteristics; urban or rural residence and geographic region; and the presence and adequacy of health insurance. This brief discussion highlights only a few of these factors.

Education level is significantly associated with overall health. Persons with less than a high school education report higher rates of physical limitations than do those with at least a high school education (Holmes et al. 2009). Education likely affects health in multiple and complex ways. One obvious mediator is health literacy, which is the extent to which persons understand and process basic health information as necessary to make appropriate health-related decisions (Berkman et al. 2011). Data pooled from more than 300 studies found that about 26 percent of patients have inadequate health literacy and another 20 percent have marginal literacy (Powers, Trinh, and Bosworth 2010). Research has linked poor health literacy to worse health, more frequent hospitalizations, lower use of preventive services, and higher mortality rates. One study among English- and Spanish-speaking diabetes patients older than age 30 found that 38 percent had inadequate health literacy and 13 percent had marginal literacy (Schillinger et al. 2002, 478). Compared to those with adequate health literacy, diabetes patients with inadequate literacy were significantly less likely to experience tight glycemic control and more likely to have poor glycemic control (as opposed to moderate control) and retinopathy. A study among Medicare beneficiaries found that persons with inadequate health literacy experienced significantly more hospitalizations than did persons with adequate health literacy, even accounting for differences in age, sex, race and ethnicity, education, income, smoking, alcohol use, chronic diseases, and self-reported physical and mental health (Baker et al. 2002b).

The contribution of persons' social network to their health is increasingly recognized. Extensive analyses from the longitudinal cohort Framingham Heart Study in Massachusetts suggest important effects of social network members on each other's health and health behaviors. Results relating to obesity were especially striking (Christakis and Fowler 2007, 370):

> A person's chances of becoming obese increased by 57% (95% confidence interval [CI], 6 to 123) if he or she had a friend who became obese in a given interval. Among pairs of adult siblings, if one sibling became obese, the chance that the other would become obese increased by 40% (95% CI, 21 to 60). If one spouse became obese, the likelihood

that the other spouse would become obese increased by 37% (95% CI, 7 to 73). These effects were not seen among neighbors in the immediate geographic location. Persons of the same sex had relatively greater influence on each other than those of the opposite sex.

The Framingham data also identified other significant social network effects. Levels of emotional health were linked to close contacts in social networks (Fowler and Christakis 2008; Hill et al. 2010; Rosenquist, Fowler, and Christakis 2011). The researchers found that emotional states—positive (happiness) and negative (including depressive symptoms)—behave similarly to infectious conditions, spreading from one social network member to another. Analyses of Framingham Heart Study data from 1971 to 2003 found that social networks have a significant influence on cigarette smoking and cessation success. Smoking behaviors spread through close and distant social contacts; groups of socially connected individuals stopped smoking together; and smokers were increasingly marginalized by their social contacts (Christakis and Fowler 2008). Social networks might also affect alcohol consumption in an active way; the persons with whom individuals interact may influence their alcohol consumption rather than simply reflect a decision to befriend persons who have similar alcohol consumption patterns (Rosenquist et al. 2010).

For those who are married or partnered, the spouse or partner is the closest relation in the social network. Considerable evidence suggests that marital status has an important influence on health and health-related behaviors. Data from the National Health Interview Survey find that married adults are more likely than unmarried adults to have better health and health-related behaviors, including lower rates of functional limitations, low back pain, headaches, serious psychological distress, smoking, and leisure time physical inactivity (Schoenborn 2004). These beneficial effects hold regardless of age, sex, race, ethnicity, education, income, and nativity. The only exceptions to this pattern are overweight and obesity, which are more common among married persons (particularly men). However, those who are married or partnered are susceptible to the "widowhood effect"; for both men and women, death of a spouse from almost any cause (e.g., cancer, cardiovascular disease, infections) heightens all-cause mortality in the bereaved spouse (Elwert and Christakis 2008b). The reasons for this phenomenon are unclear (Elwert and Christakis 2008a).

Numerous publications have documented socioeconomic disparities in health status in the United States and worldwide (Braveman and Tarimo 2002; Braveman et al. 2010). Poverty has obvious consequences. For instance, if persons can afford only substandard housing, they face heightened risks of morbidity, including respiratory infections, asthma, lead poisoning, injuries, and mental health problems (Krieger and Higgins 2002). Socioeconomically disadvantaged patients have higher mortality in the years following heart surgery than do wealthier individuals; income was significantly more predictive

than patients' race (Koch et al. 2010). Those who grow up in poverty are more likely than wealthier children to have a disability later in life (Bowen and González 2010).

Researchers have argued how poverty influences health. Given the observation that less wealthy people tend to engage in higher-risk health behaviors than do wealthier individuals, what factors are contributing to or determining their poorer health? Initiated in 1967, the Whitehall study of British civil servants produced influential insights into the effects of the socioeconomic gradient on health. The Whitehall II study, which enrolled 10,308 British civil servants in 1985 and tracked their health and health behaviors through 2004, suggested that differences in four health behaviors (smoking, alcohol consumption, diet, and physical activity) explained 72 percent of social inequalities in all-cause mortality (Stringhini et al. 2010).

People's jobs also contribute to health risks. Occupational exposures cause or heighten susceptibility to certain illnesses, such as respiratory conditions resulting from exposure to dusts, gases, or fumes. Exposure to coal dust, asbestos, silica, talc, and animal proteins can cause pulmonary fibrosis; coal dust, welding fumes, and other compounds can precipitate bronchitis; and toluene diisocyanate, chromium, grains, animal products, and cotton can induce chronic airway disease. Many such conditions are exacerbated further by smoking. Jobs also affect mental health. For instance, stresses associated with air traffic control and frontline law enforcement are well documented.

Finally, lack of health insurance affects persons' ability to receive care and thus treatment for important health conditions. For example, among persons aged 18 to 64 with diabetes mellitus, those who had no health insurance during the preceding year were six times more likely not to receive needed medical care than those who were continuously insured (47.5 percent versus 7.7 percent) (CDC 2010d). The overall percentage of Americans lacking health insurance for at least part of 2009 was 19.4 percent (58.5 million), ranging from 3.7 percent in Massachusetts to 24.6 percent in Texas[11] (Cohen, Martinez, and Ward 2010). The Patient Protection and Affordable Care Act promises nearly universal health insurance coverage, but at the time of this writing, implementation of coverage provisions is uncertain (pending Supreme Court decisions) and several years distant.

Health-Related Quality of Life

Health status and quality of life reflect patients' points of view about their overall health and how their health and other factors affect their lives. Unlike functional status, which can be measured by outsiders, only patients (or perhaps their proxies) can assess health status and quality of life. Thus, these measures provide critical insight into the outcomes of care. They also are

significant risk factors for a variety of outcomes, especially such patient-centered outcomes as future overall health status and quality of life, satisfaction with care, and mental and physical functional status.

Tools for measuring health status and quality of life may be as brief as one question: How do you rate your overall health: excellent, good, fair, or poor? Other instruments include dozens of questions on specific markers of disease, general physical capabilities, psychosocial and emotional functioning, and sense of well-being. The published literature addressing health status and quality-of-life measures has grown tremendously over the last decade, and hundreds of measurement tools have been developed for different populations and purposes.

Despite the growing number of articles addressing health status and quality of life, in many of them the exact goals of measurement are unclear. A review of 75 articles purporting to discuss quality-of-life measurement found references to 159 different instruments, with a mean of 3 instruments (range 1 to 19) per article (Gill and Feinstein 1994, 622). However, only 15 percent of the articles included a conceptual definition of quality of life, and just 17 percent stated that patients had been invited to rate their global quality of life. Thus, according to Gill and Feinstein (1994, 624), "while professing to measure quality of life, many researchers are really measuring various aspects of health status. . . . Quality of life is something that is perceived by each patient individually. The need to incorporate patients' values and preferences is what distinguishes quality of life from all other measures of health."

Eliciting patients' own values is essential to measuring quality of life because the value that each of us places on a given health state may differ from the value it holds for others. Evidence suggests that "old people tend to be health optimists, having more favorable health perceptions than their levels of physical functioning objectively allow" (Kutner et al. 1992, 534). The opinions of persons with disabilities are especially important (Dolan 1996, 559):

> . . . Those in what others may perceive to be "poor" health place a relatively high value on their own health since they have adjusted their life styles and expectations to take account of their condition. This may be particularly true of young disabled men and women, since one-quarter of this group of respondents describe their health as "poor" yet value it as "good." Conversely, young people who describe themselves as "healthy" . . . may be reluctant to value their health near the top . . . because they have high expectations about what being in the "best imaginable health state" involves. . . . More than one-fifth of respondents [without disabilities] describe their health as "good" yet value it as "poor."

Children and adolescents also have different notions of health status and quality of life than adults (Starfield et al. 1993).

Another complexity is people's perception of their health status when their quality of life is changing. Patients' perceptions of overall health—especially those of acutely ill patients—can shift over short periods, if not

daily. Therefore, questions arise about whether to capture absolute health status at one time point or its trajectory as health status worsens, improves, or remains unchanged (Covinsky et al. 1998). Gender and race or ethnicity may also confound responses to quality-of-life questions that involve lifestyle preferences (Pereira et al. 2010).

An important question is what to do when persons cannot respond for whatever reason (e.g., poor health, cognitive impairment, logistical considerations). Family members or close friends are often used as proxy respondents, but as discussed in Chapter 7, proxies may not accurately reflect patients' views. The direction of potential bias—whether proxies report better or worse views than would patients—is unclear because the research evidence is contradictory. To complicate matters, the direction of bias may vary among subpopulations (e.g., young versus elderly persons). Mode of administration (e.g., mail, telephone) also may affect responses to quality-of-life questions (Hays et al. 2009).

Some health status and quality-of-life measures relate specifically to persons with particular conditions, whereas others are generic (independent of diagnosis) (Coons et al. 2000; Hickey et al 2005; Ravens-Sieberer et al. 2006). No single method suits all research needs. Choice of approach depends on the research question. Mosteller, Ware, and Levine (1989) recommend routinely using both condition-specific and generic methods; Patrick and Deyo (1989) suggest using standardized, generic instruments with disease-specific supplements. NIH's PROMIS item bank conflates certain aspects of condition-specific considerations and overall quality of life; the PROMIS website states NIH's intention of producing "a set of rigorously designed questions about different aspects of health-related quality of life (pain, fatigue, anxiety, depression, social functioning, physical functioning, quality of sleep, etc.)" (www.nihpromis.org).

Clinical trials, including randomized clinical trials, typically measure quality of life as a critical endpoint for study participants, although different researchers use different approaches and methods could certainly be improved (Goodwin et al. 2003; Lemieux et al. 2011). In the end, it is important that persons using quality-of-life measures be clear about their goals, define their terms, and be prepared to live with some imperfection in their measures (Bergner 1989, S153–54):

> The terms quality of life, health status, and functional status are often used interchangeably and without specific definition. . . . Quality of life, just as health or illness, must be assessed specifically. . . . Somewhere in the process of deciding on the domains and choosing measures, clinical investigators often start the futile search for the measure, the gold standard that everyone will find appropriate and credible. The bitter truth is that there is no gold standard, there is unlikely ever to be one, and it is unlikely to be desirable to have one.

Patients' Attitudes and Preferences Regarding Outcomes

Patients' attitudes and preferences about outcomes, religion, cultural beliefs and behaviors, and preferences and expectations regarding care often affect their clinical outcomes and thus become putative risk factors. Some patients seek more aggressive care than others do. Aggressive interventions may delay death or impairment but may also cause treatment-related complications.

Studies suggest that about one-third of patients do not follow their physicians' recommendations, especially for preventive and outpatient care. The reasons relate largely to patients' health beliefs (especially personal views of their vulnerability and the seriousness of their condition), health-related motivations, and perceptions of the psychological and other costs of following recommendations. More than one-third of adult Americans annually seek alternatives to traditional allopathic medicine, especially for chronic conditions, but many do not tell their doctors (Eisenberg et al. 1993, 1998). Patients' goals and preferences influence not only clinical outcomes but also the costs of their care.

Culture and religion may affect compliance with prescribed therapy, diet and other daily activities, and attitudes toward health care. A prominent example is Jehovah's Witnesses' prohibition of blood transfusions regardless of clinical circumstances. Another example is pica, which is more common in the South and associated with maternal anemia and poor birth outcomes. Churches may shape community perceptions toward health and health care services by encouraging parishioners to seek care in some cases and by raising cautionary notes at other times (Markens et al. 2002). Social support from fellow parishioners, rather than specifically religious messages, may be especially influential in certain African-American communities in improving health-related behaviors and use of preventive services (Powe, Faulkenberry, and Harmond 2010; Krause, Shaw, and Liang 2011).

Patients' attitudes and preferences distill a lifetime of experiences, beliefs, goals, health status, quality of life, and understanding of prognoses and treatment options. By definition, this process is uniquely personal. Accordingly, categorizing patients' attitudes and preferences for care is complex. For example, patients differ in their preferences for end-of-life care (Barnato et al. 2007, 2009; Hanchate et al. 2009; Phelps et al. 2009), and these views may change over time as patients weigh their clinical course and personal circumstances. One study tracked preferences of 2,073 patients over two years and found that preferences for such interventions as cardiopulmonary resuscitation, artificial respiration, and tube feeding frequently changed over time (Danis et al. 1994). Changing preferences complicate efforts to include patients' attitudes among risk factors.

Some patterns of end-of-life care preferences relate systematically to certain patient attributes, such as age, race and ethnicity, marital status, decision-making capacity, and religion. In 2002, for instance, whites were

more likely than African Americans to use hospice care in the year before death (29 versus 22 percent), but both rates are relatively low and could be explained by a number of factors (Connor et al. 2008). In states with higher overall hospice use, however, racial disparities diminished. The nature of patients' underlying disease and the treatment context (e.g., inpatient, outpatient) also contribute to varying patterns of end-of-life preferences. However, demographic, sociocultural, disease, and other observable characteristics do not predict the end-of-life preferences of individual patients (Ditto and Hawkins 2005; Borreani and Miccinesi 2008). Finally, patterns of communicating care preferences vary across populations. For instance, among long-term care populations, black persons were less likely than white persons to have advance directives (Jones, Moss, and Harris-Kojetin 2011).

Evidence suggests that patients' care preferences and attitudes and providers' manner of addressing patients' desires are distinctly nonrandom. Rates of adherence to physicians' recommendations are especially poor among patients of low socioeconomic status. Highly educated, wealthier patients are more likely to obtain recommended preventive services, such as mammograms. Furthermore, patients who participate more actively in decision making about their care may experience better outcomes (Kaplan and Ware 1989). After listening to audiotaped interactions between patients and physicians during outpatient visits, Kaplan, Greenfield, and Ware (1989) rated patients' conversational styles. Patients who assumed control of conversations (e.g., asked more questions, directed the flow of the discussion and their physicians' behavior) during a baseline office visit reported missing fewer days of work, fewer health problems, fewer health-related functional limitations, and higher health status at a follow-up visit. Physiological outcome measures were also linked to patients' conversational control: At follow-up visits, patients seeking more control had lower blood glucose and blood pressure levels.

Mortality rates are a prime example of the nonrandom effect of patients' preferences on perceptions of outcomes. According to Holloway and Quill (2007, 802), "mortality has been criticized as a measure of quality for years and debates about methods of risk adjustment are almost clichéd," but neglected in these debates are concerns about "preference-sensitive care." Hospitals vary widely in the use of early DNR orders. Despite this variation, hospital mortality measures erroneously treat all deaths as medical failures. In 2007 Medicare launched Hospital Compare (www.hospitalcompare.hhs.gov; see chapters 9, 12, and 17), a website listing various performance measures, including risk-adjusted mortality rates for heart attack and congestive heart failure (CHF), for hospitals nationwide. Hospital Compare identified a Buffalo, New York, hospital as one of the 35 worst US hospitals because its July 2005–June 2006 CHF mortality rate was 4.9 percentage points higher than the national mean. The hospital reviewed the medical records of its CHF decedents and found that 11 (about 40 percent of total CHF mortalities) were in hospice or

receiving palliative care–only treatment at the patients' request (Holloway and Quill 2007, 802). Thus, these deaths were expected and did not represent the quality shortfalls targeted by the Hospital Compare initiative.

Implemented in 1991, the Patient Self-Determination Act requires health care institutions to inform patients about advance directive provisions. Nevertheless, the use of advance directives is relatively low. National surveys found that only 28 percent of home care, 65 percent of nursing home, and 88 percent of discharged hospice care patients had at least one advance directive (Jones, Moss, and Harris-Kojetin 2011). Whether clinicians follow patients' preferences is problematic. One study found that preferences for comfort care rather than life-sustaining therapies were frequently disregarded, even among persons aged 80 or older (Somogyi-Zalud et al. 2002). On the other hand, although older patients often eschew aggressive care, some do desire cardiopulmonary resuscitation and care to extend life; nonetheless, these older patients receive less aggressive life-prolonging care than do younger patients (Hamel et al. 2000). Even after controlling for patients' prognoses and preferences, older age significantly predicted decisions to withhold ventilator support, surgery, and dialysis (Hamel et al. 1999b).

Patient preferences can affect report cards on provider performance that are based on risk-adjusted outcomes. In the mid-1980s Pennsylvania began requiring hospitals to collect detailed clinical information from medical records using the MedisGroups tool. These data were used to risk-adjust hospital mortality and morbidity rates for public reporting. State analysts informed one hospital in Pennsylvania that its MedisGroups severity-adjusted mortality rate for cancer patients was much higher than expected. Even more worrisome, patients with admission scores of 0 (indicating mild if any clinical instability) had high death rates. Upon investigating, the hospital's senior oncologist found that these patients who died with scores of 0 had entered the hospital explicitly for pain control and terminal care. Physicians had talked with these patients about whether they wanted even routine testing (e.g., phlebotomy to monitor basic serum chemistries), but the patients had requested comfort measures only. Thus, the standard blood tests used by the MedisGroups measures (e.g., serum sodium, potassium, hematocrit, white blood cell count) were not performed. Without measurement, no clinical abnormalities were identified, hence producing severity scores of 0. In this circumstance, the 0 scores represented the desire of terminally ill patients to maximize their comfort at the end of life, not the absence of severe disease.

Additional Issues

A few loose ends relating to risk factors remain: the role of processes of care, the nature of interventions, and random chance.

Role of Processes of Care

As noted throughout this book, risk factors aim to capture attributes patients bring to health care encounters that might affect their risks for particular outcomes. Typically, the goal of risk-adjusting outcomes is to isolate the effectiveness or quality of care from the patient-related risk factors. In this context, adjusting for factors related to quality could mask quality shortfalls and thus compromise the ability of risk-adjusted outcomes information to provide meaningful insight about quality.

Certain types of procedures and other processes of care (e.g., administration of certain medications) are sometimes used to treat complications resulting from substandard care. Risk adjustment methods that adjust for these procedures or care processes could confound efforts to isolate effectiveness or quality by masking quality problems and thus producing misleading results. Furthermore, adjusting for procedures or processes of care raises the potential for manipulation by providing incentives for practitioners to perform these procedures to bolster their risk-adjusted outcomes results. The potential for such gaming is especially worrisome when the procedure (or process of care) is useful for multiple purposes or discretionary (i.e., not dictated by scientific evidence), leaving considerable room for clinicians to make individual judgments about their use.

However, in designing their risk adjuster for congenital heart disease surgical mortality in pediatric populations (Chapter 13), Jenkins and colleagues (2002) argued that little if any discretion exists in the operations performed for various heart defects. Working extensively with panels of physicians, they repeatedly returned to specific operations as crystallizing patients' clinical risk factors for surgical outcomes. Their Risk Adjustment for Congenital Heart Surgery (RACHS-1) measure used expert consensus and empirical methods to determine relative risks of in-hospital death associated with specific surgical procedures and other clinical characteristics. In this highly technical area, Jenkins and collaborators (2002) believe that surgeons cannot manipulate the operative approach for pediatric heart surgery patients simply to game or manipulate their risk-adjusted outcomes. As described in Chapter 13, however, this position has become controversial. Other experts have suggested that there are opportunities to manipulate assignment of procedure codes, thus compromising the integrity of RACHS-1 results.

Other investigators have developed risk adjusters based on pharmacy claims, using prescription information to indicate patients' burden of disease (Johnson, Hornbrook, and Nichols 1994; Gilmer et al. 2001; Fishman et al. 2003; Sloan et al. 2003; Kuhlthau et al. 2005; Maciejewski et al. 2005; Powers et al. 2005; Stam, van Vliet, and van de Ven 2010). For example, the developers of the Chronic Illness and Disability Payment System and the Diagnostic Cost Group model released versions based on pharmacy claims: Medicaid Rx and RxGroups, respectively (Gilmer et al. 2001; Zhao et al. 2001,

2005; Sales et al. 2003). Not surprisingly, these models significantly predict annual costs of care. However, pharmacy claims reflect not only patients' conditions but also practice patterns. In many instances, the decision to prescribe a drug or a specific type of drug is discretionary and highly variable across physicians. Depending on their use, these risk adjusters could have counterproductive effects; according to Gilmer and colleagues (2001, 1201), "Pharmacy-based risk adjustment may reward those plans and providers that prescribe drugs liberally, and punish those that have adopted more conservative prescribing practices."

Nature of the Intervention

Beyond risks arising from patient attributes, the nature of the intervention being studied is important; the treatment may present its own risks. For example, major surgery requiring general anesthesia typically raises more immediate risks than does surgery performed under local or spinal anesthesia. In some instances, complications (e.g., idiosyncratic but deadly reactions to anesthetic agents) result from the intervention itself rather than from the patient's disease or risk factors. In many circumstances, surgery is contraindicated for patients with extensive disease in major organ systems, such as the lungs and heart. Thus, the decision to forgo surgical intervention because of serious coexisting diseases could confound observational comparisons of surgical and medical therapies.

Similarly, certain medical therapies pose immediate inherent risks. The express goal of most chemotherapeutic regimens is to destroy malignant cells; this process may result in well-anticipated and defined physiological derangements. Chemotherapy side effects may be particularly problematic for patients with certain comorbid illnesses (e.g., cardiomyopathy, renal failure). Therefore, among chemotherapy patients, certain acute physiological derangements (e.g., low white blood cell counts, high uric acid or potassium levels because of tumor lysis) may indicate the presumed effectiveness of treatment. Persistence of such abnormalities, however, generates concern. The nature and timing of these "intentional" abnormalities resulting from treatment must be considered in risk assessment.

Random Chance

Despite comprehensive efforts to capture risk factors, some important attributes inevitably elude detection or quantification. As noted earlier, no risk adjustment method includes all pertinent risk factors, and even the predictions of reasonably comprehensive risk adjusters are imperfect. Although risk adjusters sometimes predict with good statistical assurance patients' outcomes across populations, outcomes for individual patients are sometimes unexpected. These unanticipated outcomes can result from random chance or "noise" (Luft and Romano 1993, 336):

Whenever the focus is on outlier events, the group of outliers will be composed of "normals" who experienced bad outcomes by chance and true "abnormals" who actually had a high risk of bad outcomes. This problem arises particularly when patient populations are small and poor outcomes are rare.

For example, randomness can cause the majority of the differences in mortality rates across hospitals (Park et al. 1990). Concerns about interpreting findings based on small sample sizes and randomness are particularly pressing when looking at clinician-specific risk-adjusted outcomes. As described further in Chapter 12, when sample sizes are small, the confidence intervals around risk-adjusted outcomes estimates are generally large, raising questions about the rigor of these findings. Most clinicians have argued, with some statistical legitimacy (see Chapter 17), that their individual patient panel sizes are too small to examine clinician-level risk-adjusted outcomes, especially for relatively rare events (e.g., deaths).

Notes

1. Exceptions are sperm and egg cells, which contain 23 chromosomes each. When sperm cells fertilize egg cells, the father's 23 chromosomes link with the mother's counterpart chromosomes, giving the new organism the full complement of 46 chromosomes.
2. CC, also known as Copy Cat, lives with a Texas family at the time of this writing. She was ten years old as of late 2011. She bore three healthy kittens and has led a normal life, apart from all the publicity.
3. Differences in questions and response categories contributed to some discrepancies across survey findings about levels of functional and other health declines among older individuals (Freedman et al. 2004; National Research Council 2009).
4. A study by the US General Accounting Office (1992) (now the US Government Accountability Office) found that 60 percent of clinical trials for new drugs underrepresented female subjects. Of 53 drug trials studied, only 25 (47 percent) specifically assessed whether men and women responded differently to the medication. Female representation was particularly poor in trials involving new cardiovascular drugs; in 7 of 13 cardiovascular medication trials examined, the proportion of female subjects was more than 20 percent below the percentage of women in the population with the disease in question. Similar underrepresentation of women in federally funded research studies prompted the creation of the NIH Office of Research on Women's Health.
5. The six conditions were acute myocardial infarction (AMI), congestive heart failure (CHF), obstructive airway disease, gastrointestinal

hemorrhage, pneumonia, and stroke. Patients admitted as interhospital transfers were excluded. After controlling for risk factors, the study found that across all conditions except AMI and CHF, men died in hospital significantly more often than women did.

6. The Thirteenth Amendment abolished the three-fifths rule, and American Indians were fully counted starting in 1924 (Williams 1999).
7. SOFA variables are a respiratory indicator (PaO2/FiO2 mmHg), coagulation (platelets 103/mm2), bilirubin, hypotension, Glasgow Coma Score, and creatinine.
8. For example, if a surgical patient dies, a surgeon could inflate the patient's condition to make it appear that death was almost inevitable when in fact the patient's death was due to error on the surgeon's part (and thus the surgeon's rating would not reflect poor performance).
9. The need for total assistance with bathing produced an adjusted odds ratio of death (95 percent CI) of 6.69 (2.89–15.49) for stroke patients and 4.98 (2.74–9.08) for pneumonia patients (Davis et al. 1995, 913, 915).
10. I have placed lesbian, gay, bisexual, and transgender health issues here, under the header of psychological, cognitive, and emotional functioning. This topic needs to be in this chapter because of the small but compelling body of evidence about health risks associated with minority sexual identity. I debated, however, where best to place this critical attribute, recognizing the precise fit among chapter sections is unclear.
11. These numbers reflect the percentage of residents in different states who said they were uninsured at the time they responded to the National Health Interview Survey.

CHAPTER

4

WINDOWS OF OBSERVATION

Amy K. Rosen and Ann M. Borzecki

The importance of specific risk factors depends on the answer to the second essential question presented at the beginning of Chapter 2: over what time frame? Is the outcome of interest (e.g., mortality) imminent, within minutes or hours, or years away? For example, current blood glucose levels indicate whether patients with diabetes are at immediate risk of ketoacidosis or hyperglycemic coma. But for more distant outcomes, such as retinopathy, peripheral vascular disease, or lower-extremity amputation, indicators of longer-term glycemic control (e.g., hemoglobin A1c) are more relevant risk factors. Therefore, when developing or assessing risk adjustment methods, we must specify a time window. Most risk adjusters focus on a specific time frame, such as one year of health care utilization or a hospitalization period. This chapter examines the implications of different time windows for conceptualizing risk and identifying risk factors.

Conceptual Framework

Windows of observation are the time frames or events circumscribing the outcomes of interest. The definition of the window of observation typically derives from the targeted outcome. For example, the window of observation for hospital mortality is the time from admission to discharge (alive or dead) or some short, fixed time frame triggered by hospitalization (e.g., 30 days after admission).

Fixed time windows (e.g., a certain number of days or months) help equalize comparison of outcomes across pertinent units of observation (such as hospitals) with varying practice patterns. For example, more than 20 years ago, Jencks, Williams, and Kay (1988) found that Massachusetts hospitals had higher in-hospital mortality rates for Medicare beneficiaries than did California hospitals. At the time, California hospitals had dramatically shorter lengths of stay than did Massachusetts facilities. The longer the time window, the more deaths were detected. When Jencks and colleagues fixed the window of observation from admission to 30 days hence, differences in hospital death rates between the two states disappeared.

As suggested at the beginning of the chapter, some risk factors are more or less useful depending on the window of observation. For example, extremely high or low blood pressure is an acute physiological derangement that portends imminent death; however, extreme hypertension or hypotension can result from acute pathophysiological processes (e.g., idiosyncratic drug reaction, severe bleeding, shock from trauma) that, if successfully treated, may have few long-term consequences. While extreme blood pressure readings heighten risk of death within hours or days, they may not predict longer-term outcomes, such as mortality a year or two hence. In contrast, a lower level of persistent hypertension may present no short-term risk but may raise the risk of future events, such as stroke or congestive heart failure (CHF). The usefulness of sociodemographic, economic, lifestyle, and cultural risk factors (see Chapter 3) also depends on the window of observation. Some of these non-clinical factors, such as Jehovah's Witnesses' prohibition of blood transfusions, have implications for short-term outcomes, but most other social factors, such as smoking and poverty, are linked to longer-term outcomes.

The window of observation relates directly to the outcome of interest. However, the time window from which risk factor information is drawn depends on the purpose of the analysis and on the database used. For instance, if the purpose of a hospital mortality analysis is to isolate facilities with potential quality-of-care problems, using risk factors or conditions present on admission (POA) is essential. They must be differentiated from complications that develop during the hospitalization. If they are not, quality-of-care problems could be masked. For example, low blood pressure or hypotension present on admission strongly predicts death and is not related to care received at the admitting hospital. If risk factors are not differentiated from complications, low blood pressure that occurs several days after admission would similarly be associated with an increased risk of death but could have resulted from failure to carefully monitor the patient and recognize sepsis or septic shock. Therefore, inclusion of these later blood pressure readings as risk adjusters in mortality models would mask problems at hospitals that fail to adequately monitor their pneumonia patients. The recent introduction of POA codes for the principal and secondary diagnoses should help distinguish conditions present on admission from those that develop during the hospitalization (Coffey, Milenkovic, and Andrews 2006; see Chapter 5).

Characteristics of databases also could affect the time frame from which analysts draw data. Information from before a hospital admission could provide important insight into risk factors for hospitalization outcomes. For instance, hospitals often perform routine preoperative testing up to 30 days before they admit patients for elective surgery. If a different facility, not the admitting hospital, performs this testing, the admitting hospital's medical records may contain little information from the preadmission evaluation

(whether this information is incorporated into the admitting hospital's medical records depends on a number of factors). If analysts do not look at longitudinal administrative records (claims or encounter records from months to a year prior to admission; see Chapter 5), they will not gather and consider information about chronic conditions that could be important for risk-adjusting hospitalization outcomes. As described in chapters 3, 9, 12, and 17, the Medicare Hospital Compare risk adjusters for acute myocardial infarction (AMI), CHF, and pneumonia are collected through this longitudinal approach.

Some risk factors may be important to prediction of outcomes for a current period but not for future periods. For example, algorithms analysts use to predict costs over one year of care often come in two versions: a *concurrent model*, which predicts costs in the year the predictor variables (typically diagnoses) occur, and a *prospective model*, which predicts costs in future years, generally the year following that from which the diagnoses are drawn. Sometimes, significant predictors of costs in the current year are not predictive of the following year's costs. A clear example is appendicitis. Appendicitis increases costs in the current year because of the surgery and hospitalization required to treat the condition. But once the appendix is removed, appendicitis will not recur; the condition will never increase future costs.

Risk factor information from surveys (see Chapter 7) may carry a time window. For example, functional status questionnaires often ask people about functional abilities within specific time frames, such as the prior 30 days. When using such time-sensitive information, analysts may find it important to link the date of the survey with the date of the index event (e.g., a hospitalization) or outcome of interest (e.g., mortality). As noted in Chapter 3, the best predictor of future functional status is previous functional status (e.g., some designated baseline), but analysts must take care not to create a tautology. For example, using current self-reported functional status to predict functioning in the short term may be inappropriate given the short time frame. The appropriate timing for measuring baseline and future functional status depends on the analytic questions and context of the study.

Finally, in conjunction with the outcome of interest, the purpose of the study helps define the appropriate time window. For example, to profile providers with respect to processes of care or outcomes (e.g., costs), one might look at an episode of care for managing a specific disease or condition. Here, the unit of observation is the episode. The start and end points of the episode will vary with the condition being examined and the purpose of the study. For example, if the study focuses on profiling providers with respect to costs, the episode might be defined as a fixed time frame, such as a year (e.g., in the case of management of a chronic condition, such as diabetes) or a discrete episode of acute illness (e.g., in the case of management of an acute condition, such as myocardial infarction). If the goal is to predict the resource needs of a given

population, a fixed time frame, such as a year, is the most appropriate unit of analysis.

The method of risk adjustment used also depends on the study purpose. When examining provider performance with respect to processes or outcomes, one might use risk adjustment to improve the clinical homogeneity of the episode by accounting for differences in patient characteristics, such as the severity of illness or conditions that are likely to contribute to variation in intermediate and long-term outcomes. When examining costs, analysts need to incorporate risk variables, such as patients' comorbidities, that account for the type and cost of procedures and treatments delivered during the time frame of interest. To predict population resource needs, analysts need to account for the cumulative effect of patients' various conditions within the specified time frame. A variety of patient attributes might be important risk factors, including demographic characteristics, diagnoses, health-related behaviors, functional status, and self-reported global health status.

Therefore, when designing a risk adjustment strategy, one must answer questions about time window: risk of what outcome over what time frame? The window of observation has important implications for which risk factors are most relevant as analysts develop their conceptual model of risk (see Chapter 8). Depending on the purpose, database, setting, and context, window of observation options are almost infinitely varied. Here, we describe in greater depth three common time windows—acute care hospitalization; fixed, longer-term windows; and episodes of care.

Acute Care Hospitalization

Acute care hospitalizations are perhaps the most obvious window of observation and the focus of extensive study. For almost four decades, researchers have used increasingly sophisticated risk adjustment approaches to examine costs, lengths of stay, and outcomes potentially related to quality of care. Therefore, many risk adjusters exist for evaluating outcomes of acute care hospitalizations. For several decades, Diagnosis-Related Groups (DRGs) were the best known, most widely used method for risk-adjusting prospective Medicare payments for acute care hospitalizations (Fetter et al. 1980).

As noted earlier, defining when windows or episodes of care start and end is essential, both for identifying relevant risk adjusters and identifying outcomes. The beginning and ending of acute care hospitalizations seem obvious; the window starts when patients are admitted and closes at discharge. However, operationally these time points are often unclear. Especially over the last several years, administrative and local environmental concerns have sometimes blurred the exact time of admission. Because of complex insurance coverage issues, some hospitals keep certain patients in observation beds

rather than formally admit them to the facility; these patients may eventually be admitted if they fail to recover quickly. Emergency department (ED) practices also vary across hospitals. In overcrowded EDs and full hospitals, patients may wait many hours before being admitted for an acute illness. Time to admission also varies by location; patients presenting to EDs in crowded inner-city teaching hospitals are typically more likely to experience delays than are patients presenting to EDs in suburban facilities.

Determining the end point of acute hospitalizations may also be complicated. One concern is patient transfer, not only to other acute care hospitals but also to settings such as acute rehabilitation hospitals and nursing homes. For example, in the past, post-stroke rehabilitation usually began during lengthy acute care hospitalizations. Now patients generally are transferred as soon as possible to post–acute or sub–acute care facilities. Transfer practices vary by hospital type (e.g., academic teaching versus community facilities), health care system, and geography and depend on the services available in the local community.

The impetus for speedy discharges also varies by hospital ownership. For instance, hospitals run by the Veterans Health Administration (VHA) have traditionally allowed longer stays than have private facilities. One study compared in-hospital mortality among ICU patients admitted to a VHA hospital to in-hospital mortality among patients admitted to 27 private-sector hospitals in 1994 and 1995 (Kaboli et al. 2001). Average length of stay was much higher at the VHA hospital than at the private hospitals (28.3 versus 11.3 days), and unadjusted mortality was also higher (14.5 versus 12.0 percent). These differences disappeared after adjusting for illness severity. More important, a higher proportion of VHA patient deaths occurred after 21 days in hospital. After Kaboli and colleagues (2001) applied proportional hazards regression models (see Chapter 10), censoring patients at hospital discharge, the VHA patients had a lower risk of death (hazard ratio 0.70, $p < 0.001$).

With the recent surge of interest in reducing readmissions, the impetus behind speedy discharges may decrease as hospital managers and policymakers become increasingly aware of the critical importance of adequate discharge planning in reducing readmissions (Jencks, Williams, and Coleman 2009). Nonetheless, factors affecting the timing of hospital admission and discharge still vary widely across hospitals, potentially biasing comparisons using hospitalization as the window of observation. Perceptions of outcomes may shift dramatically after varying time windows are taken into account. To deal with differing discharge practices, analysts often define hospitalization windows using fixed periods following admission (e.g., 30, 90, or 180 days), which are less dependent on hospital practices. For example, the Centers for Medicare & Medicaid Services (CMS) reports 30-day risk-adjusted mortality and readmission on its Hospital Compare website (www.hospitalcompare.hhs.gov).

While use of a 30-day window seems conceptually superior to an "in-hospital" window for outcomes such as death, in-hospital measures are more straightforward to derive and readily available, requiring only hospital discharge abstracts. Of note, the Agency for Healthcare Research and Quality (AHRQ) developed the Inpatient Quality Indicators (IQIs) and the Patient Safety Indicators (PSIs) (www.qualityindicators.ahrq.gov), which use hospital discharge abstract data (see Chapter 5) to screen for quality-of-care problems occurring in the inpatient setting. IQIs include mortality indicators: in-hospital death rates associated with specific procedures or conditions that have shown provider variation and for which evidence exists that high mortality may be associated with poorer care (AHRQ 2006). IQIs use the All Patient Refined Diagnosis-Related Groups (APR-DRGs) for risk adjustment, while PSIs use aggregated DRGs (i.e., DRGs with and without complications are collapsed into a single DRG) and a modified version of a comorbidity index developed by Elixhauser and colleagues in an attempt to distinguish comorbidities present on admission from in-hospital complications (Elixhauser et al. 1998; AHRQ 2007). Both IQIs and PSIs are increasingly being used for hospital profiling and public reporting (Remus and Fraser 2004). The National Quality Forum (NQF) endorsed several of the mortality IQIs and PSIs as hospital performance measures, and CMS added selected mortality IQIs and PSIs to its Hospital Compare website.

As described in Chapter 3, important risk factors for acute care hospitalizations include acute clinical stability and acute attributes of principal diagnoses and comorbid conditions. However, depending on the outcome of interest, chronic conditions can also significantly predict acute care results, and physical functional status is a critical risk factor for in-hospital mortality for many conditions. Therefore, longer-term risk factors can significantly predict outcomes even within short time windows.

Fixed, Longer-Term Windows

Another way of specifying time windows is to look at the calendar. If one aims to capture outcomes for populations in a system of care regardless of where, when, or how services are provided, it makes sense to focus on a fixed block of time. In today's health policy context, for example, total costs expended in caring for a particular population (e.g., enrollees of a health plan) are often an important outcome. For this accounting, analysts typically choose one year (12 months based on calendar or fiscal year).

Most of the early interest in one-year windows focused on prior costs. However, risk quantified over a year is now used to adjust capitation payments to managed care plans, profile expenditures by provider, and assist allocation of resources. CMS adopted the Hierarchical Condition Category Model, which accounts for diagnoses associated with both acute hospitalizations and

outpatient services, to determine next year's prospective capitation payments for managed care organization enrollees (Pope et al. 2004).

Yearlong risk adjustment also has expanded to other applications, such as identifying persons with particular medical problems for disease management programs or potentially high-cost persons for case management. As noted earlier, the model that analysts use depends on the time frame of the outcome. Concurrent (same-year) models are often used to understand how disease burden affects historical costs or to produce provider profiles. Prospective models use information from the current or base year to predict future (next year's) costs or to identify next year's high-cost patients.

While the beginning and end points of 12-month periods can be clearly specified, members of the target population may be difficult to identify. When populations are defined by insurance status, analysts generally use plan enrollment dates to identify those in the target population—but insurance coverage may fluctuate over the course of a year. Fluctuating insurance status (i.e., churning) is particularly common among Medicaid recipients, whose coverage depends largely on their income. For example, a study found that among 4.7 million California Medicaid enrollees, 62 percent had at least one coverage interruption during the period from 1998 through 2002[1] (Bindman, Chattopadhyay, and Auerback 2008). In these instances, analysts often look at each month of coverage and reweight information for such individuals by how many twelfths of a year they were covered. For example, the Diagnostic Cost Group (DCG) model annualizes the costs for individuals with less than one full year of entitlement in the prospective year on the basis of their observed cost per month; in analyses, these data are treated as "fractional observations" (Ellis and Ash 1995).[2] Thus, a person with costs of $10,000 who has entitlement through the third month has annualized costs of $30,000. Similarly, using the Chronic Disease Payment System (CDPS) to analyze residuals from unweighted regression models for Medicaid beneficiaries, Kronick and colleagues (2000) found that one method of reducing the influence of shorter entitlement periods was to weight each observation by $(1 - 0.067 \times [12 - \text{number of eligible months}])$.

Longer windows of observation are typically used to examine outcomes for populations, regardless of the setting of care. Risk factors are often drawn from various sources, including acute care hospitalizations and outpatient visits. As noted earlier, acute physiological stability is probably not a significant risk factor for longer-term outcomes, but virtually all other dimensions of risk are potential candidates depending on the outcome of interest, the target population, and the purpose of the evaluation.

Risk Adjusters for Longer-Term Windows

Various existing risk adjustment tools focus on predicting total costs (or expenditures) for one year of care for insured populations. Taking one of these

methods "off the shelf" may be appropriate in some settings, but risk adjustment methods developed for one population or purpose may not perform as well for other purposes or populations with different characteristics. Researchers sometimes supplement these off-the-shelf models with disease-specific indicators relevant to their study populations, as well as sociodemographic variables (e.g., marital status, race/ethnicity, socioeconomic status), to improve the models' predictive ability (Ettner et al. 1999; Ash et al. 2000; Rosen et al. 2002). Pharmacy data are sometimes used as proxies for disease burden in predicting health care resource use for a 12-month period (Fishman and Shay 1999; Gilmer et al. 2001; Zhao et al. 2001; Sales et al. 2003). Models combining pharmacy claims and inpatient data have performed significantly better than either model alone and "provided a more complete picture of the distribution of illness in the population" (Zhao et al. 2001, 2005).

Although most risk adjusters for one-year outcomes have focused on predicting resource use, important exceptions include the comorbidity index developed by Charlson and colleagues (1987), the Index of Coexistent Disease (Greenfield et al. 1993), and the comorbidity index developed by Elixhauser and colleagues (1998). The first two indexes were derived from medical records of hospitalized patients to predict risks of one-year mortality and one-year health-related quality of life, respectively. Although primarily used to predict clinical outcomes within a 12-month period, the indices may predict outcomes over longer time frames, such as five-year survival (Krousel-Wood, Abdoh, and Re 1996). The Charlson Comorbidity Index may also predict short-term mortality (Poses et al. 1996) and is frequently used in studies of acute hospitalization outcomes. While the Charlson Comorbidity Index has been adapted to use administrative data (Deyo, Cherkin, and Ciol 1992), Elixhauser and colleagues (1998) developed their comorbidity index specifically to predict in-hospital outcomes using inpatient administrative data and to be more comprehensive and applicable across a wide range of diseases than existing indices. It specifies 30 sets of individual ICD-9-CM diagnosis codes (secondary codes), or comorbidities. The word *index* is somewhat of a misnomer, in that conditions are accounted for individually rather than combined into an index. A modified version of the index with 27 comorbidities, developed as part of AHRQ's Healthcare Cost and Utilization Project (HCUP 2011; see Chapter 5) and used to risk-adjust AHRQ's PSIs, is available for download in SAS code at www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp.

Episodes of Care

An *episode of care* is "a series of temporally contiguous health care services related to treatment of a given spell of illness or provided in response to a specific request by the patient or other relevant entity" (Hornbrook, Hurtado,

and Johnson 1985). An episode framework relates health care inputs (e.g., specific events, primary care and specialist services, time required to produce particular outcomes) to health care outputs, such as duration and course of illness. Many persons, especially elderly patients, have multiple concurrent chronic conditions that wax and wane and putative episodes of care with fuzzy boundaries. Erecting strict perimeters specifying the start and finish of a health problem defies this reality. Nevertheless, an episode approach facilitates analysis of both processes and outcomes of care because it groups together some or all of the services related to the care of a patient's acute or chronic conditions over a specified period. Episodes of care are a popular management tool, and interest in using episodes of care as a basis for "bundled" payments has grown. They also are considered a framework for performance measurement (i.e., for profiling providers on the basis of their resource use/costs, quality, and efficiency [defined as measures of the cost of care associated with a specified level of quality]) (NQF 2009). These new ways of applying episodes of care present additional challenges in risk adjustment (discussed at the end of this chapter).

Episodes usually require risk adjustment before they can be compared meaningfully. Patients in the same type of episode should be reasonably clinically homogeneous with respect to the specific health problem addressed by the episode and therefore would be expected to experience similar patterns of resource use (Rosen and Mayer-Oakes 1999). Although determining which risk factors to incorporate into an episode is complicated, several characteristics of episode approaches suggest which risk factors should be considered (Hornbrook, Hurtado, and Johnson 1985). The first consideration is the target outcome of the episode. For example, analyses examining patient health status at the ends of episodes should focus on patient-specific information, such as pre-episode risk (e.g., patients' baseline health status, including severity of illness and comorbidities) and patients' preferences and health-related behaviors, whereas episode-of-care analyses evaluating resource utilization may incorporate not only patient-specific information but also provider- and system-level factors, such as access issues and wait times. If the purpose of the analysis is to characterize the nature of the episode (i.e., acute or chronic) rather than the patient, patient risk factors such as patients' psychosocial functioning and health-related behaviors may be less important than they would be in other contexts.

A second consideration is episode duration. Episodes involve one or more encounters and vary in length. The beginning and end points of episodes are defined differently in different contexts. For example, episodes may commence with the patient's initial contact with the health care system or when a diagnosis is confirmed. Episodes may end within predetermined intervals (e.g., 30 days after the index visit for an acute episode or within 12 months for a chronic episode) or when the condition is cured or resolves.

"Clean periods" are a third consideration (Rosen and Mayer-Oakes 1999). Clean periods are time intervals in which patients did not receive services and are often used to define the end of an episode; new episodes begin when services resume after the clean period. In the clean period approach, end points may represent varying outcomes; for example, episodes end when patients no longer seek care for whatever reason, when they leave a particular system of care, or when treatment is no longer necessary. As time windows lengthen, patient-level risk factors and comorbidities become more important. As in other settings, the more time that elapses, the more likely it is that factors outside providers' control (e.g., specific processes of care) will influence outcomes. The potential to confound quality of care with risk factors also increases as time frames lengthen. Short episodes, on the other hand, typically provide "snapshots" of care and can be tightly defined to isolate the effect of specific conditions; herein, the primary focus is on the patient's acute or chronic condition and the services associated with it.

A fourth consideration is the course of the episode: the rate of progression or natural history of the targeted condition. Some conditions have rapid onsets and progress quickly, while others advance or recede slowly. Variability in the rate of a condition's progression is also important. Patients with identical diagnoses may experience widely varying courses of illness, perhaps due to inherent physiological variations or other factors, such as comorbid conditions or physical frailty. It is therefore important to capture disease progression through the course of the episode and the variability in the rate of a condition's progression.

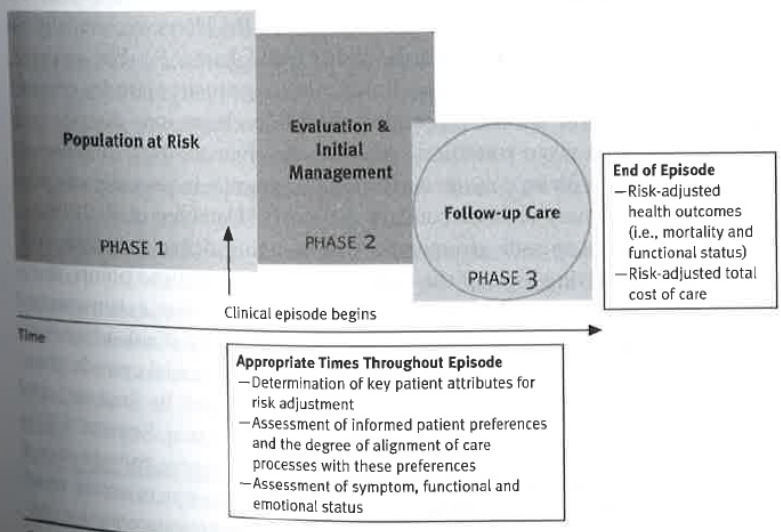
A few additional considerations may not require risk adjustment. For example, to specify episodes, analysts must first identify the clinical trigger that starts the episode, such as a procedure or visit for a certain diagnosis. That trigger could be defined in detailed clinical terms, potentially making risk adjustment unnecessary.

Until recently, the episode framework was considered more suitable for acute than chronic diseases. In theory, acute diseases start and stop at definite points, and the course of acute disease is well defined. In contrast, chronic diseases often have indeterminate beginnings and endings, and the course of chronic disease is typically 90 days or more, often spanning years or lifetimes (Hornbrook, Hurtado, and Johnson 1985). Analysts have tried to create more "acute-like" episodes from chronic diseases through a variety of strategies. One strategy, using diabetes as an example, involves dividing a chronic episode into sub-episodes or phases, such as a diagnostic phase (when diabetes is first recognized), a maintenance phase (routine management of glycemic control), an acute flare-up phase (when metabolic disturbances such as ketoacidosis develop), and a chronic complication phase (for conditions resulting from end organ damage, such as foot ulcers) (Rosen and Mayer-Oakes 1998, 1999).

Each phase may have its own set of risk factors. Another approach divides chronic disease episodes into fixed-length intervals, such as one-year periods.

The preferred use of episodes for acute rather than chronic diseases has changed over time, in part due to the development of a "generic episode of care" model by the NQF (2009). This model was designed to assess efficiency of care (association of cost and quality). The episode framework was selected primarily because it provides a longitudinal perspective on how care is coordinated across multiple settings and at critical transition points. The generic model is applicable to multiple types of episodes and a broad set of acute and chronic diseases. It has been applied to AMI, substance abuse, cancer, diabetes, and low back pain (NQF 2008a, 2008b; Rosen and Makay 2009) and can be used to track the population at risk, evaluation and initial management, and follow-up care that occur over the course of an episode of care (NQF 2009). Exhibit 4.1 depicts this model and its different phases, including where in the episode the NQF committee thought risk adjustment should be considered.

The following two examples illustrate differences in the use of risk adjustment in this generic model for the same disease and between diseases. For AMI, an acute disease, NQF developed different trajectories that have varying outcomes and patterns of care. Patients in trajectory 1, who are relatively healthy at the time of their initial AMI, are expected to return to their normal activities following recovery if care is delivered effectively. Risk adjustment is not considered an important component of this model. Patients



Source: Reproduced with permission of National Quality Forum ©2009.

EXHIBIT 4.1 Generic Episode of Care: National Quality Forum

in trajectory 2, who have AMI and other serious underlying conditions (i.e., comorbidities), are expected to recover more slowly and function at a lower level than they did pre-AMI. The complexity of patients' disease burden needs to be accounted for in all phases of this trajectory.

In contrast, for low back pain (often a chronic condition), the diagnosis and initial management phase is most important for determining patients' clinical conditions and establishing the presence of comorbidities, such as depression, that may affect outcomes (NQF 2009). Also important are the shared decision making and informed choice that occur between patients and their providers in the evaluation and initial management phase.

Implementing Episode Algorithms

In general, administrative data such as claims or encounter databases containing diagnosis and procedure codes, codes for each service, and service dates (see Chapter 5) can easily be used to operationalize episodes of care. Using these data, analysts can array services in chronologic order. Various commercial software products called "episode groupers" use retrospective computerized algorithms to cluster services into discrete episodes of care to support provider profiling, quality assessment, outcomes measurement, case identification for disease management, and payment. The risk factors that the groupers prioritize as most important include inpatient and outpatient primary and secondary diagnosis codes and patients' sociodemographic characteristics. Some groupers also include pharmacy claims and procedural information.

Episode grouping software products have good face validity (MedPAC 2005, 2006). (Face validity is discussed in Chapter 9.) However, virtually no studies have examined their criterion validity (see Chapter 9). For example, do the groupers accurately identify clinically homogeneous episodes of care? More fundamentally, do they appropriately group claims into discrete episodes of care for a given condition? Additionally, their accuracy in grouping claims associated with an episode is affected by variation in provider or coder assignment of primary versus secondary diagnoses (Damberg et al. 2009).

Commercial episode groupers vary in how they define the target conditions, the underlying logic of the episodes (e.g., start and end points, duration of clean periods, types of claim records included, whether a claims record is assigned to more than one episode), and the complexity of risk adjustment (Rosen and Mayer-Oakes 1999). The two leading commercial episode groupers are Episode Treatment Groups (ETGs), developed by Ingenix, and Medical Episode Grouper (MEG), developed by Thomson Reuters. ETGs are conceptually similar to DRGs, except they classify an entire episode of care rather than a hospitalization. Both ETGs and MEG attempt to create more clinically homogeneous groups with respect to expected resource use by classifying episodes by severity.

ETG version 7.0 may assign a given claims record to multiple episodes. The software includes 524 base ETGs, each of which has its own clean period marked by the absence of treatment. Base ETGs have no predetermined length except for chronic episodes, which are fixed at a 12-month period (MaCurdy et al. 2008). ETGs further classify 129 of the base ETGs into up to four severity levels, yielding a total of 679 distinct episode classifications (MaCurdy, Kerwin, and Theobald 2009). The grouper also assigns a severity score to each episode on the basis of age, sex, complications, comorbidities, and interactions among complications and among comorbidities (Ingenix 2008).

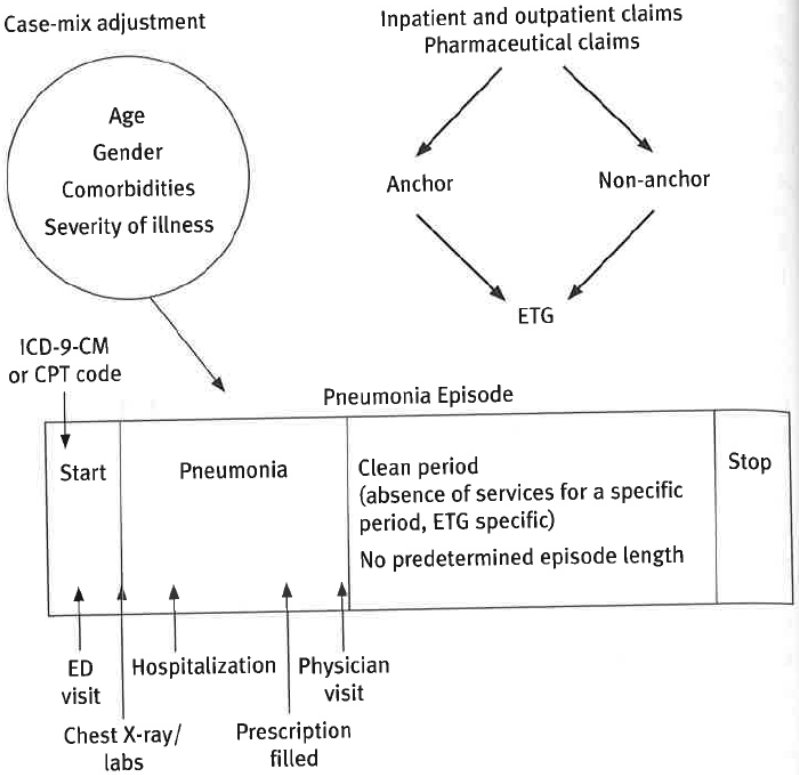
In contrast, MEG version 7.1 assigns each claim to one episode and assigns each episode to one of 560 base episode groups (MaCurdy et al. 2008). These groups are then classified into severity levels on the basis of organ system, etiology, and severity of illness using disease staging (Gonnella, Hornbrook, and Louis 1984), stage 1 representing the best level of health status and stage 4 representing the worst (death). MEG supplements risk adjustment using the DCG model to estimate a patient's expected costs on the basis of the patient's disease complexity and derive a relative risk score on the basis of the medical problems treated in a specified period.

Exhibit 4.2 depicts how the ETGs classify an episode of pneumonia. Because ETGs and MEG are fairly similar conceptually and operationally, we first describe how the ETG software groups the claims and then point out differences in the MEG software.

Consider this typical clinical scenario: A patient presents to the ED with shortness of breath, fever, and cough. The patient undergoes a chest X-ray and laboratory tests and is admitted to the acute medical ward with a diagnosis of bacterial pneumonia. After three days, the patient is discharged and given a prescription for antibiotics (a seven-day course). The patient obtains the medication, adheres to the prescribed regimen, and follows up with the primary care physician several days later. This series of events generates several claims.

ETGs first group claims into record type (i.e., management, surgery, ancillary [laboratory or X-ray], facility, or pharmacy). The software then identifies anchor records (i.e., records that represent direct clinician services) and non-anchor records (i.e., ancillary or pharmacy records, which represent services that are secondary to the direct evaluation and management services) (MaCurdy et al. 2008). MEG also groups types of records similar to those that ETGs group. In both ETG and MEG software, an episode is initiated by a provider office visit or a hospitalization. They both group claims on the basis of associated codes and use ICD-9-CM diagnosis codes and CPT-4 (Current Procedural Terminology) codes for outpatient procedures. However, for inpatient procedures, ETGs use Revenue Center Codes while MEG

EXHIBIT 4.2 Episode Treatment Groups, Pneumonia Example



uses ICD-9-CM procedure codes. An episode is complete when no treatment is provided for a specified period (i.e., the clean period).

In the pneumonia example in Exhibit 4.2, the ED visit initiates an episode. In the ETG software, the ED visit, hospitalization, and follow-up physician visit are all considered anchor records, while the claims for the laboratory tests and chest X-ray (ancillary records) and the prescription fill (pharmacy record) are non-anchor records. These records are categorized into an ETG base class "437400" and one of four possible full ETGs for bacterial lung infections on the basis of codes indicating the presence of complications or comorbidities. In the MEG software, these same records are assigned a group code "510" (bacterial pneumonia) and a severity level (stage) of 1 to 4 (subcategories exist within severity levels 2 and 3). Both ETGs and MEG look for related claims occurring up to 60 days after the last physician office visit. When none is found, they end the episode on the date of that visit.

Challenges for Risk Adjustment

Use of an episode framework may present important challenges, including questions about performing risk adjustment. These issues are especially pressing when episodes are used to profile provider performance and bundle health care payments.

Performance Profiling

Although episodes are an appropriate time window for profiling providers, this approach raises a number of issues. Episode definition (whether single or multiple conditions are included in the episode) is particularly critical. Because patients may have more than one condition (i.e., multiple co-occurring conditions or comorbidities), a "single-condition-focused episode" definition would not account for other conditions being treated concurrently. For example, Hussey and colleagues (2009) found that Medicare beneficiaries "had an average of eight or more episodes during a year, some of which were for interrelated conditions: beneficiaries who had an AMI also had hypertension (63 percent), heart failure (54 percent), or diabetes (35 percent) episodes." Consequently, risk adjustment would be inadequate, and provider profiling on outcomes most frequently used (resource use or costs, quality, and efficiency) would be unreliable if it focused on a single condition; it would capture only a narrow view of patients' clinical management.

On the other hand, creating a "multi-condition-focused episode" definition (i.e., one that folds multiple conditions into a single episode) would provide a broader perspective on patients' care and be more reliable. However, this approach also has drawbacks in that it might create more heterogeneous episodes despite risk-adjusting outcomes (e.g., overall resource utilization) for comorbid conditions. In addition, sorting out which process(es) of care "belongs" with which condition(s) is difficult and further complicates the reliability of provider profiling. Thus, given the challenge of adjusting for the complexity of patients' multiple comorbidities for profiling purposes, episode groupers need to account for not only the different services related to each condition but also their associated conditions.

A few empirical studies demonstrate challenges in episode-based profiling. Hussey and colleagues (2009) found considerable variation in standardized payments for treatment of specific conditions, suggesting that this variability might be explained, in part, by the heterogeneity of a patient's condition (e.g., severe versus mild pneumonia), which is not adequately captured by claims data. Similarly, MaCurdy and colleagues (2008) used both ETGs and MEG on Medicare claims data to profile providers on the basis of costs. They found a high degree of skewness (particularly in the top 5 percent) and variation in the distribution of costs within episode types, even though, as noted earlier, each of the groupers incorporates severity levels (i.e., ETG severity and MEG staging categories) and supplemental modules for risk

adjustment. For example, for episodes of bacterial lung infections, costs from the 10th to 95th percentile ranged from $58 to $9,307 in ETGs and from $40 to $11,885 in MEG (coefficients of variation ranged from 1.30 to 2.18). The authors concluded that "using one of the prominent commercial grouping products without more refined risk or severity adjustments could readily yield a performance system that improperly rewards providers for factors or behaviors beyond their control" (MaCurdy et al. 2008). Further work suggests that, even when risk adjustment is used, providers may be unfairly penalized because episode construction may not adequately reflect the full course of care and may capture high-cost institutional treatment as a separate episode from low-cost follow-up noninstitutional treatment (MaCurdy et al. 2010).

Another issue related to episode-based profiling is the reliability of profile scores—that is, the proportion of variability in a measure due to real differences in performance (the signal-to-noise ratio). Adams and colleagues (2010) found that 22 percent of physicians were misclassified on the basis of cost profile scores created using ETGs; reliability across specialties ranged from 0.05 to 0.79. MaCurdy and collaborators (2010) also examined the reliability of profile scores by examining stability of scores over time. For both ETGs and MEG, year-to-year correlations were modest (0.46 to 0.60) when all episodes were examined and lower when disease-specific episodes were examined (0.41 to 0.45). Because a physician's cost profile (i.e., the sum of the observed costs divided by the expected costs for all assigned episodes) relies partly on risk adjustment (the expected costs), unless risk adjustment accounts for multiple co-occurring conditions and other factors, the reliability of provider profiles is likely to be low. Interestingly, to the contrary, Thomas (2006) found that with the severity stratification methodology used in ETGs, the ETG software produced scores that were accurate enough to profile providers on the basis of cost efficiency reliably.

Episode-Based Bundled Payments

Use and endorsement of an episode approach for payment reform has gained momentum in the past few years (Davis 2007; IOM 2007; MedPAC 2008; Bigalke 2010). Episodes are attractive as a basis for payment because groups of services and costs for a clinical condition can be bundled into one episode, encouraging better coordination of care across providers and settings and thus better outcomes and lower costs (Bigalke 2010). The assumption behind episode-bundled payments is that providers would be appropriately incentivized to improve quality of care because multiple providers would be jointly accountable for the total cost of care through shared payment (Keckley, Underwood, and Frink 2009). CMS is planning to implement bundled payments to hospitals in three phases. In phase one, which is set to begin in October 2014 (fiscal year 2015), bundled payments will be made for admissions for conditions that account for the top

20 percent of post–acute care spending (i.e., conditions that account for the top 20 percent of spending on services provided during the first 30 days following an acute care inpatient hospital stay) (Murer 2009; Bigalke 2010). In other words, payment for post–acute care services no longer will be separate from payment for the inpatient stay; one payment will cover both. Several examples of episode-based bundled payment systems are being tested or implemented (Keckley, Underwood, and Frink 2009), including PROMETHEUS Payment, Geisinger Health System's payment for cardiac care episodes, and the Medicare Physician Hospital Collaboration demonstration (Hussey et al. 2009).

PROMETHEUS Payment, being pilot-tested by CMS, is an episode-based payment model that defines global case rates for given conditions on the basis of clinical practice guidelines, risk stratification, and an allowance for potentially avoidable complications (Rosenthal 2008). This model produces an "evidence-informed case rate" (ECR). ECRs aim to quantify all the resources needed to provide optimal care (i.e., purportedly evidence based) to a patient with a given condition. The risk adjustment component accounts for patient characteristics, including demographics, comorbid conditions, severity of illness, evidence-informed procedures performed, and evidence-informed pharmacy use (www.hci3.org). Comorbidities are addressed by increasing the base amount negotiated by a provider to account for the additional resources needed to treat the patient. If the comorbidities change the basic reason for treatment, the ECR may be divided into two or more ECRs according to the number of comorbidities. However, this subdivision depends on whether the patient has distinct conditions or distinct processes of care to which evidence-based guidelines can be applied.

Although in theory ECRs are risk adjusted and designed as homogenous clinical grouping systems, highly complex patient cases are unlikely to be included in the PROMETHEUS Payment model due to the complexity of accounting for comorbidities within a single episode and aligning the episode with clinical practice guidelines. Thus, patients whose care requires application of more than one set of clinical guidelines will potentially be eliminated (Prometheus Payment Inc. 2006). Finally, because expected payments are based on average expected patient costs within an episode, risk adjustment may inadequately account for those who are severely ill.

Conclusion

Different windows of observation affect the range of risk factors that should be considered in developing risk adjustment methods. Risk factors vary in importance, depending on the time frame. The time interval also has significant implications for the meaning of risk-adjusted outcomes information,

especially for calculating the "algebra of effectiveness": the extent to which risk factors—as opposed to health care interventions, the quality of that care, and random chance (see Chapter 2)—contribute to outcomes.

Notes

1. The absence of a monthly eligibility code following the presence of an eligibility code defined an interruption of coverage.
2. For newborns, an annual period is used because of timing differences between birth dates and contract eligibility dates.

CHAPTER

5

CODED DATA FROM ADMINISTRATIVE SOURCES

Lisa I. Iezzoni

Administrative data result from running the health care system: processing health insurance enrollment applications, certifying coverage; approving expenditures; adjudicating reimbursement amounts; paying claims; tracking service utilization; and monitoring providers' costs, practice patterns, and performance. Administrative data are not produced explicitly to examine the health or health care of populations, but they are frequently used for these purposes for the following reasons:

- They include large numbers of people, sometimes entire populations (e.g., all persons hospitalized in a state).
- They represent care practiced throughout the community rather than in specialized settings.
- Linkage of individuals' administrative records tracks persons' health over time and across settings of care.
- Federal regulations standardizing data content and formats make certain administrative data items comparable across public and private health care delivery systems.
- Although individual identifiers are removed from the data to protect confidentiality, large numbers of people further hide patients' identities.
- Administrative data already exist, are relatively inexpensive to acquire, and are digital.

Administrative data have provided profound insights into US health care practices. Almost four decades ago, Wennberg and Gittlesohn (1973) used hospital discharge data to expose wide variations in the use of medical interventions across small geographical areas with ostensibly similar populations. In 1989, Section 6103 of the Omnibus Budget Reconciliation Act (P.L. 101-239), which created the federal Agency for Health Care Policy and Research (AHCPR) (later renamed the Agency for Healthcare Research and Quality [AHRQ]), mandated use of large administrative databases to examine the "outcomes, effectiveness, and appropriateness" of health care services. AHCPR's inaugural flagship projects, the Patient Outcomes Research Teams, used administrative data (Clancy and Eisenberg 1997).

Despite their usefulness, administrative data have inherited two significant limitations from the stubbornly fragmented American health care delivery system. First, most person-level databases include only the claims and encounter records of people who have public or private health insurance and thus contain no data on uninsured individuals, estimated at 50.7 million (16.7 percent) in 2009 (DeNavas-Walt, Proctor, and Smith 2010, 22). The Patient Protection and Affordable Care Act (ACA) aims to ensure that most Americans have health insurance, but those nearly universal coverage provisions will not take effect until 2014. In addition, high turnover, which is common among Medicaid recipients and private employment-based health insurance enrollees (due to a continuously changing work force), impedes efforts to create longitudinal, population-based databases. How ACA provisions will affect this turnover or "churn" is unclear, especially given new provisions relating to Medicaid recipients (see discussion later in this chapter).

Second, for claims and encounter records to be entered into databases, the service provided must be covered by health insurance; however, many important services, particularly for chronic conditions, are not covered, especially by Medicare and private health plans. Examples include long-term rehabilitation therapies, many assistive technologies, and personal care assistance. Therefore, even though people may spend thousands of dollars annually out of pocket for health-related services, these costs are not captured in administrative databases because no claim was submitted for this care. Thus, the services and medical conditions indicated by claims and encounter files may not fully reflect health care experiences and needs.

Administrative data have other limitations. Most important, their primary clinical insight comes from diagnoses coded with questionable accuracy, completeness, clinical scope, and meaningfulness (Iezzoni 2002). In particular, administrative data files do not contain information about patients' preferences, such as DNR and comfort-measures-only requests. These gaps can severely compromise the utility of administrative data for producing meaningful information about mortality rates, for example (Chapter 3). Furthermore, administrative databases are typically large and messy and require maintenance by sophisticated programmers who understand the management of complex data sets. Although timeliness is improving, many administrative data files remain somewhat out of date; data may not be available for up to two years following an event.

Despite these problems, administrative data provide important information about health services utilization, expenditures, selected clinical outcomes, and quality of care. Managers, policymakers, and researchers rely heavily on these data sets to examine important outcomes of care, and many risk adjustment methods were designed specifically for administrative databases. As described in Chapter 6, the promise that electronic clinical information will be readily available for download from providers' information

systems remains unfulfilled, although recent policy changes aim to speed adoption of this technology. Therefore, administrative data files likely will continue to be the source of critical information about health care costs and quality into the foreseeable future.

Overview of Administrative Databases

Administrative data are the by-product of operating and overseeing the health care system. In some administrative databases, the unit of observation is a specific service (Iezzoni 2002). The best example is data on acute care hospitalizations. Most states systematically collect electronic information about each discharge in the state submitted by hospitals, typically using the standardized Uniform Bill (UB) format (currently UB-04, see discussion later in this chapter), and sometimes collect various additional data. With certain exceptions, individual patients do not have unique statewide identification numbers, preventing tracking of an individual's admissions across hospitals in these state databases. Furthermore, these databases generally cannot link with other settings of care.

AHRQ's Healthcare Cost and Utilization Project (HCUP 2011) has compiled states' hospitalization information since 1988. HCUP now includes 40 states and represents more than 90 percent of discharges from US community hospitals.[1] HCUP State Inpatient Databases (SID) are available for download from AHRQ's HCUP website after users sign a Data Use Agreement in which they agree not to try to identify individual patients and to use the data only for research and statistical purposes. Exhibit 5.1 shows selected SID variables among more than 100 contained in the data set.

In health insurance databases, individuals are usually the unit of observation and receive unique identifiers, enabling their services to be linked across care covered by the insurer.[2] Most health insurance databases contain two types of files: enrollment files, indicating eligibility for the health plan and basic demographic information (Iezzoni 2002), and claims (submitted by fee-for-service [FFS] providers for payment) or encounter records (submitted by providers within managed care organizations), representing individual services or sets of services. We focus here on person-level administrative data produced by public (e.g., Medicare, Medicaid, TRICARE, US Department of Veterans Affairs [VA]) and private health insurers. Persons outside the organizations that produce these databases, such as the Centers for Medicare & Medicaid Services (CMS) and the VA, can sometimes gain access to the data after meeting specific confidentiality and security requirements.

Most clinical insight from administrative data comes from records of individual services. Major administrative information throughout the health care system must comply with transaction standards specified pursuant to

EXHIBIT 5.1 Selected Variables Contained in the State Inpatient Databases (SID), AHRQ's Healthcare Cost and Utilization Project*

Variable	Variable Name	Variable Description
Variables About Admission		
ADATE	Admission date	
DDATE	Discharge date	
ASCHED	Admission scheduled	Indicates whether admission was scheduled
ADSOURCE	Admission source	1 = emergency department; 2 = another hospital; 3 = another health care facility, including long-term care; 4 = court, law enforcement; 5 = routine
ATYPE	Admission type	1 = emergency; 2 = urgent; 3 = elective; 4 = newborn; 5 = trauma center; 6 = other
READMIT	Readmission	State-specific indicator, if provided, that hospitalization is a readmission
LOS	Length of stay	
DIED	Died in hospital	
DISP	Disposition of patient	1 = routine; 2 = short-term hospital; 3 = skilled nursing facility; 4 = intermediate care facility; 5 = another type of facility; 6 = home health care; 7 = against medical advice; 20 = died
	Payers	Primary and secondary payers
Variables About Patient, Diagnoses, Care, and Costs		
AGE	Age in years	Age in years at admission date
AGEDAY	Age in days	Age in days when age is less than one year

EXHIBIT 5.1
(Continued)

Variable	Variable Name	Variable Description
AGEMONTH	Age in months	Age in months when age is less than 11 years
FEMALE	Sex	
RACE	Race	1 = white; 2 = black; 3 = Hispanic; 4 = Asian or Pacific Islander; 5 = Native American; 6 = other
HISPANIC X	Hispanic ethnicity	
PrimLang	Primary language	Primary language of patient, coded using state-specific codes
MARITALSTATUSUB04	Marital status	Marital status as coded on UB-04: I = single; M = married; A = common law; B = registered domestic partner; S = separated; X = legally separated; D = divorced; W = widowed; U = unmarried (single or divorced or widowed)
HOMELESS	Patient is homeless	Provided by selected states
ZIP INC QRTL	Zip code median income	Quartile classification of the estimated median income for the patient's zip code
	Patient location	More than a dozen different variables indicating urban-rural location
APGAR1	1-minute APGAR score	APGAR score measured 1 minute after birth A = activity and muscle tone; P = pulse, heart rate; G = grimace response; A = appearance, skin coloration; R = respiration, breathing rate and effort

(Continued on next page)

EXHIBIT 5.1
(Continued)

Variable	Variable Name	Variable Description
APGAR5	5-minute APGAR score	APGAR score measured 5 minutes after birth
BWT	Birth weight	Birth weight in grams
DNR	DNR indicator	Indicator of presence of do-not-resuscitate order, provided by few states
	Detailed charges	Different variables for detailed and line item charges received from hospital
DSNDX	Maximum number of diagnoses	Maximum number of diagnosis codes provided by the source, range 0–30
DSNPR	Maximum number of procedures	Maximum number of procedure codes provided by the source, range 0–30
DXATADMITn	Diagnosis present on admission	Coded for each diagnosis; indicates whether the condition was present on admission to hospital; started in 1998
DXn	Diagnoses	ICD-9-CM principal and secondary diagnosis codes; the first-listed code indicates the principal diagnosis
PRn	Procedures	ICD-9-CM procedure codes; the first-listed code indicates the principal procedure
PRDATEn	Procedure date	Procedure date masked to protect patient identity
	Days	Various variables indicating the number of days spent in specific settings in the hospital, such as different types of ICUs
	Utilization flags	Indicators of utilization of approximately 30 specific types of services, such as renal dialysis, echocardiography, and various imaging studies

EXHIBIT 5.1 (Continued)

Variable	Variable Name	Variable Description
Variables About Hospital and Physicians		
AHAID	AHA hospital identifier	American Hospital Association hospital identification number
HOSPST	Hospital state	Hospital state postal code
	Physician number	Masked number for up to four physicians
	Physician specialty	Specialty for up to two physicians
Variables Produced from Classification Software		
	DRGs	DRGs produced using various versions of the software
ADRG	All Patient Refined DRG	APR-DRG from 3M Health Information Services software
ADRGRISKMORTALITY	APR-DRG mortality risk	1 = minor likelihood of dying; 2 = moderate likelihood of dying; 3 = major likelihood of dying; 4 = extreme likelihood of dying, as computed by APR-DRGs
ADRGSEV	APR-DRG severity level	APR-DRG complexity subclass 0 = newborn DRGs; 1 = minor loss of function (includes cases with no comorbidity or complications); 2 = moderate loss of function; 3 = major loss of function; 4 = extreme loss of function
	Comorbidity variables	More than 30 comorbidity indicators generated by AHRQ's Clinical Classification Software (CCS)

*SID contains more than 100 variables, some taken directly from hospital discharge abstracts and others computed using software or algorithms. This table shows only selected data elements from SID.

Source: Data from HCUP (2008).

Title II of the 1996 Health Insurance Portability and Accountability Act (HIPAA; P.L. 104-191). HIPAA mandates standardized content, formats, and code sets of various computerized records, making their specifications comparable across public and private payers.

For purposes of comparison, administrative files may be sorted into two types: standard files, typically governed by HIPAA requirements, and clinically enriched files, generally specified for particular programmatic purposes (such as capturing information on nursing home or home health agency patients). Standard administrative files contain less clinical information than do the clinically enriched files.

Standard Administrative Data

Helpful clinical information in standard administrative files includes

- diagnoses coded using a clinically modified version of the *International Classification of Diseases* (ICD), currently the ninth revision (*International Classification of Diseases, Ninth Revision, Clinical Modification* [ICD-9-CM]; a transition to the tenth version [ICD-10-CM] is scheduled for 2013 but might be delayed until 2014);
- procedures and services from claims submitted by institutional providers, such as hospitals, coded using ICD-9-CM (a transition to the ICD-10-PCS [Procedure Coding System] is scheduled for 2013); individual physician services, coded using the American Medical Association's Current Procedural Terminology (currently CPT-4); and nonphysician services not included in CPT-4 (including administration of certain intravenous medications and the purchase/rental of durable medical equipment), coded using the Healthcare Common Procedure Coding System (HCPCS); and
- prescription drugs coded using the US Food and Drug Administration's National Drug Code Directory (NDC).

Various rules govern assignment of codes and thus the clinical content of standard administrative records. Many of these rules trace back to the Uniform Hospital Discharge Data Set (UHDDS) specified in 1972 by the National Committee on Vital and Health Statistics (NCVHS), then an advisory body to the US Department of Health, Education, and Welfare (DHEW) and now to the successor agency, the US Department of Health and Human Services (HHS). NCVHS (1980) aimed to create a uniform but minimum (i.e., not burdensome) data set to facilitate investigation of the costs and quality of short-term hospital services at local and national levels. In 1974, following initial testing, DHEW required submission of UHDDS data for all acute hospital discharges paid through Medicare and Medicaid. In 1979, a committee convened by DHEW revised the directions for completing the UHDDS

while retaining its original 14 data elements. Instructions of special interest, which have changed little over the intervening years, include designation of diagnoses and procedures, as follows:

- A maximum of five diagnoses should be recorded, including "all diagnoses that affect the current hospital stay." In particular, the "principal diagnosis is designated and defined as: the condition established after study to be chiefly responsible for occasioning the admission of the patient to the hospital for care"; "other diagnoses to be designated and defined as associated with the current hospital stay are: all conditions that coexist at the time of admission, that develop subsequently, or that affect the treatment received and/or the length of stay. Diagnoses that relate to an earlier episode which have no bearing on the current hospital stay, are to be excluded" (NCVHS 1980, 12). One critical change since 1980 is the much larger number of diagnosis coding slots available, which varies somewhat in different settings (e.g., some states allow 25 diagnosis codes on their hospital discharge abstracts).
- All "class 1, 2, and 3" procedures (i.e., "surgery" and "significant procedures") should be recorded along with the dates they were performed. One procedure is to be designated the principal procedure using the following criteria: "(1) The principal procedure is one that was performed for definitive treatment rather than one performed for diagnostic or exploratory purposes, or was necessary to take care of a complication. (2) The principal procedure is that procedure most related to the principal diagnosis" (NCVHS 1980, 12).

UHDDS rules relating to definition of principal diagnosis and additional diagnoses still apply (NCHS 2010b, 86, 89).

Hospitals initially submitted UHDDS data on paper forms. Until 1992, Medicare allowed only five coding slots for diagnoses, one principal plus four secondary diagnoses. Medicare forms offered three coding slots for procedures. With electronic formats, most recently UB-04 introduced by Medicare in March 2007, came increasing numbers of diagnosis and procedure coding slots. UB-04 has a principal diagnosis slot, an admitting diagnosis slot, and spaces for up to 17 other diagnoses and three external cause of injury codes (CMS 2006a). UB-04 also has spaces for a principal procedure, plus date, and for five other procedures and their dates.

Standard administrative data represent aggregations of claims, like UB-04, submitted by FFS providers when billing for their services. Rules governing submission of encounter records by managed care organizations (MCOs) vary across different payers. For example, before 1997 Medicare MCOs were not required to submit encounter records to the Health Care Financing Administration (HCFA; now CMS). When the 1997 Balanced

Budget Act mandated health-based payments to Medicare MCOs, they were required to submit encounter records that provided diagnostic information for risk adjustment. CMS requirements for MCO reporting are complicated and have changed over the years (Asper and Mann 2011). MCOs can choose to submit encounter records for different purposes, including (in select situations) direct payment to the providers submitting the claim. Hospice claims must always be submitted to Medicare, even by MCOs.

The result is that MCO claims or encounter records are not routinely included in Medicare utilization files, and researchers must typically eliminate MCO enrollees from studies that use Medicare administrative data. In 2009, 11.1 million (26.1 percent) of Medicare's 42.4 million beneficiaries were enrolled in managed care plans (Asper and Mann 2011, 2). Excluding MCO enrollees therefore means a significant fraction of Medicare beneficiaries are not included in administrative data–based studies. The rules for submission of claims or encounter records by Medicaid MCOs differ from those for Medicare MCOs, and private MCOs that serve non-Medicare and non-Medicaid populations also have varying requirements for submission of claims or encounter records.

Clinically Enriched Data

Because the data are required by law or regulation to be submitted to the government, some clinically detailed information is technically administrative. Given this extensive clinical content, I view these data as clinically enriched—the second broad category of administrative data. CMS (and sometimes state Medicaid programs and private payers) has mandated gathering extensive information about patients' functional status and disabilities in long-term or post–acute care settings (Iezzoni 2010). As described in Chapter 16, to collect these data, nursing homes must use the Minimum Data Set (MDS), home health agencies must use the Outcome and Assessment Information Set (OASIS), and inpatient rehabilitation facilities (IRFs) must use the Patient Assessment Instrument (IRF-PAI). CMS uses MDS, OASIS, and IRF-PAI data to set prospective payment amounts and in some instances to assess quality of care. Although each method collects similar types of information, the tools have important differences, and various efforts over the years to streamline and homogenize these data-gathering approaches have not yet succeeded (Iezzoni 2010).

Sources of Administrative Data

The major sources of administrative data are federal health programs, primarily Medicare and the VA (TRICARE and Indian Health Service data are used less commonly by investigators); Medicaid, a joint federal and state program; state governments (e.g., hospital discharge abstracts compiled by HCUP); and

private insurance companies. Currently, the best source of information about Medicare and Medicaid data sets is the Research Data Assistance Center (ResDAC), a CMS contractor at the University of Minnesota (www.resdac.org). ResDAC can also assist with the Data Use Agreements required to obtain these data sets.

> . . . ResDAC has a 5-year, $5 million contract from CMS as the exclusive provider of complete, adjudicated Medicare and Medicaid claims data to researchers. Under the contract, ResDAC provides free expert assistance to researchers. The cost price of data acquisition for researchers, however, can reach $50,000 and beyond depending on the research question. This price reflects fixed labor and data infrastructure costs distributed over a relatively small user base. Competition could increase access and reduce cost. (Conway and VanLare 2010, 1007)

In the future, more sources of administrative data may become available. For example, the Community Health Data Initiative is a new public–private endeavor that intends to help individuals and communities understand their health and health care performance, aiming to provide data support for improvement efforts. Under that program, HHS is providing to the public, without cost, a collection of Community Health Data Sets created from a variety of federal government sources. These data sets contain "a wealth of easily accessible, standardized, structured, downloadable data on health care delivery, health, and determinants of health performance at the national, state, and county levels, as well as by age, sex, race/ethnicity, and income (when available). The initiative offers hundreds of measures of health care quality, cost, access, and public health through a single access point" (Conway and VanLare 2010, 1008).

Medicare

Congress established Medicare in 1965 as Title XVIII of the Social Security Act to provide benefits for eligible beneficiaries aged 65 or older beginning in July 1966. The 1972 amendments extended coverage to younger persons with disabilities and persons with end-stage renal disease. Traditional Medicare includes two parts: Part A, hospital insurance (coverage for services provided in institutional settings, including hospitals, skilled nursing facilities, and hospices, and for some home health services), and Part B, supplemental medical insurance (coverage for physician services, outpatient hospital services, certain medical equipment, and other services). Enrollment in Part B is voluntary. Since the late 1990s, the Medicare+Choice program (also called Part C) has allowed beneficiaries living in certain regions to enroll in private managed care plans. The Medicare Prescription Drug, Improvement, and Modernization Act of 2003 created Medicare Part D, which in January 2006 began supporting prescription medication coverage for Medicare beneficiaries. In fiscal year 2008, Medicare had more than 44 million beneficiaries,

generating total payments of about $428 billion—15 percent of the total federal budget (CMS 2009).

While administering Medicare, CMS processes massive quantities of beneficiary, institutional, billing, and other administrative data. Medicare's administrative records are compiled into huge, computerized data files reflecting CMS's three primary functions: (1) enrolling and tracking beneficiaries, each of whom receives a unique identification number; (2) designating and monitoring providers and institutions approved to accept Medicare payments; and (3) reimbursing health plans and providers for services. In 1991, HCFA partitioned the country into nine distinct processing sectors, each with a designated contractor or "host" producing a Common Working File (CWF). Every beneficiary was uniquely assigned to a CWF host, which maintained information on that beneficiary's eligibility status and Part A and Part B utilization records (the Health Insurance Master Records). All institutional provider and physician/supplier claims submitted daily by the nine CWF hosts were entered weekly into the National Claims History 100 percent Nearline File, which contained all claims submitted for Medicare beneficiaries, including adjustments, interim claims, and denials.

Medicare's complicated and enormous claims processing infrastructure was costly and cumbersome, attracting notice when the 2003 Medicare Modernization Act was crafted. Section 911 of the act required CMS to replace its claims processing system by 2011. As of 2006, CMS contractors were processing approximately 1 billion Medicare claims annually from more than 1 million health care providers (CMS 2006b). The claims processing infrastructure included 23 fiscal intermediaries and 17 carriers, all processing FFS claims. (A separate, equally complicated processing infrastructure managed Part D billing.) Fiscal intermediaries handled the claims for facilities, including hospitals and skilled nursing facilities. The carriers processed claims for Medicare Part B, such as physician, laboratory, and other service claims. Four fiscal intermediaries focused on home health and hospice claims, while four carriers served as regional claim processors for durable medical equipment suppliers (e.g., claims for wheelchairs, prosthetics, and supplies). In addition to processing claims, these contractors enrolled health care providers in Medicare and taught them about Medicare billing requirements, handled claims appeals, answered questions from beneficiaries and providers, and investigated allegations of fraud and abuse.

Section 911 of the 2003 Medicare Modernization Act required CMS to replace this fiscal intermediary and carrier structure with competitively selected Medicare Administrative Contractors (MACs) by 2011. The new structure aims not only to streamline claims processing but also to reduce the potential for fraud and abuse. CMS found that 7.8 percent of 2007 Medicare reimbursements did not comply with coverage, billing, payment, or coding rules, and these lapses generated $24.1 billion in under- or overpayments (CMS

2009, 2010b). Included in the new claims processing infrastructure is an "alphabet soup" of entities/functions in addition to the MACs, including

- Medical Review (MR), which is performed by MACs before or after claims are paid (not all claims are reviewed; MRs target known problematic areas; for example, all new claims submitted by providers with track records of submitting incorrect claims may be reviewed);
- National Correct Coding Initiative (NCCI) edits, performed automatically before claims are paid;
- Medically Unlikely Edits (MUEs), performed automatically before Part B claims are paid;
- Comprehensive Error Rate Testing (CERT) edits, conducted following claims payment (in 2009, 95,000 randomly sampled claims were reviewed in depth, including some medical record reviews);
- Recovery Audit Contractors (RACs), made permanent by the Tax Relief and Health Care Act of 2006 (Section 6411 of the ACA expanded the recovery audit role to Medicare Parts C and D and Medicaid; RAC audits may include medical record reviews); and
- Zone Program Integrity Contractors (ZPICs), which are responsible for identifying and investigating suspected fraud cases and taking corrective actions (CMS 2010b).

This complex Medicare claims processing infrastructure is still maturing, but it seems more likely to eliminate administrative records that result from fraudulent practices than did the previous system. Whether it is simpler than the prior system is unclear.

As noted earlier, the most efficient strategy for obtaining Medicare data is to contact the ResDAC. Three broad categories of data are available, classified by the extent to which the data reveal individual identities. Various federal provisions, including the HIPAA Privacy Rule,[3] the Privacy Act of 1974,[4] and the Freedom of Information Act, determine the availability of these data sets, as follows:

- Limited Data Set (LDS): LDS data contain health information on individual beneficiaries but exclude the direct identifiers specified in the federal Privacy Rule. Despite not including these identifiers, LDS data are considered identifiable data sets, subjecting these files to the Privacy Act. To obtain LDS data, researchers must demonstrate that their planned use of the data meets the disclosure provisions for research purposes as defined by both the Privacy Rule and the Privacy Act. The research purpose ultimately must aim to improve the care provided to Medicare beneficiaries or administration of the Medicare program (ResDAC 2012). Researchers must pay to obtain these data.

- Research Identifiable Files (RIFs): RIFs contain person-specific information on Medicare beneficiaries and providers, including individual identifiers of the beneficiary or physician (e.g., birthdate, age, race, sex, residence) (ResDAC 2012). RIF data are therefore subject to the federal Privacy Act, Freedom of Information Act, and other governmental rules and regulations. CMS follows strict security safeguards to protect individual privacy. The CMS Privacy Board reviews requests for these data and determines whether applicants are eligible to obtain the files. Researchers must pay for the costs of producing the RIF data sets.
- Non-Identifiable Data Files: CMS strips the Non-Identifiable Data Files of all information that could identify individuals. These data sets generally contain aggregate information about Medicare beneficiaries or providers.

Exhibit 5.2 provides examples of Medicare files available to researchers.

Section 723 of the 2003 Medicare Modernization Act required HHS to make Medicare data about beneficiaries with chronic conditions readily available to researchers. The resultant database, the Chronic Condition Warehouse (CCW; ccwdata.org), selected its longitudinal cohort using the 5 percent national Medicare sample from 1999 to 2004; all beneficiaries in that cohort were tracked continually over time. From 2005 forward, CCW contains information for 100 percent of enrolled Medicare beneficiaries with one or more of 21 chronic conditions (e.g., acute myocardial infarction, Alzheimer's disease, breast cancer, depression, diabetes, glaucoma, heart failure, hip fracture, osteoporosis, stroke) identified from diagnosis and procedure data on Medicare claims.[5] Most important, all information gathered using the MDS, OASIS, and IRF-PAI about these beneficiaries is merged with the CCW data. The 2008 CCW contained 2.4 million Medicare beneficiaries: 2 million persons aged 65 or older and 400,000 younger individuals eligible because of disability (Iezzoni 2010). Although this database offers a rich source of functional information, all these data are derived during provision of specific Medicare-covered services, raising the potential for bias relating to differences in service availability or use by individual patients (AHRQ 2010b).

Medicaid

Medicaid is a joint state and federal program enacted as Title XIX of the Social Security Act Amendments of 1965. Although details vary by state, Medicaid covers three broad areas: (1) health insurance for low-income families and persons with disabilities, (2) long-term care for older persons and those with disabilities who meet income standards, and (3) supplemental coverage for low-income Medicare beneficiaries for services not covered by Medicare and costs of Medicare premiums and deductibles (HHS 2000).

EXHIBIT 5.2 Examples of Medicare Data Files*

Limited Data Set (LDS)	
LDS Denominator Files	LDS Denominator Files are available as 5 percent and 100 percent samples (1999–2009). Denominator Files contain demographic and enrollment information about each beneficiary enrolled in Medicare during the calendar year, including MCO participation.
MedPAR	National MedPAR (1997–2009) contains inpatient hospital final action stay records. Long-Term Care Hospital (LTCH) MedPAR is a subset of the National MedPAR and contains only LTCH discharges (2004–2008). Skilled Nursing Facility (SNF) MedPAR (2002–2008) contains SNF final action stay records. Each MedPAR record summarizes all services provided to a beneficiary from the time of admission through discharge and thus could contain information from more than one claim. In LDS MedPAR, identifiers for the individual beneficiary and the institution to which the beneficiary is admitted do not provide more detailed geographic information than state of residence/location.
LDS Standard Analytical Files (SAFs)	LDS SAFs contain information collected by Medicare to pay for health care services provided to a Medicare beneficiary in a calendar year (2000–2009). The record unit is the claim (some care episodes may have more than one claim). LDS SAFs do not contain specific service dates, presenting all dates as a quarter and year. Age is presented as a five-year age range; unique physician identification numbers (UPINs) are encrypted, and the county is the lowest level of geographic identification. SAFs are available as follows: inpatient (5 and 100 percent samples); SNF (5 and 100 percent); outpatient (5 and 100 percent); home health agency (5 and 100 percent); hospice (5 and 100 percent); carrier (formerly called physician/supplier Part B) (5 percent); and durable medical equipment (5 percent).
LDS DATE SAF	SAFs containing actual dates of service are available starting with 2009 data.
Research Identifiable Files (RIFs)	
Vital Status Files	Two versions of Vital Status Files are available: files with and without beneficiaries' names and addresses. The version with names and addresses is often used by researchers who wish to contact specific Medicare beneficiaries for their research. Vital Status Files include demographic information about each beneficiary ever entitled to Medicare, such as unique identifier, state and county codes, zip code, date of birth, date of death, sex, and race.

(Continued on next page)

EXHIBIT 5.2 (Continued)

Beneficiary Summary Files	Beneficiary Summary Files contain demographic and enrollment information about each beneficiary enrolled in Medicare during a calendar year (data are "frozen" in March of the following calendar year). Information includes beneficiary unique identifier, state and county codes, zip code, date of birth, date of death, sex, race, age, monthly entitlement indicators (for Medicare Parts A, B, C, and D), reasons for entitlement, state buy-in indicators, and monthly managed care indicators (yes/no). As of 2006, these files added variables specific to enrollment in Part D. Other variables include a derived race/ethnicity code, an indicator for Other Credible Drug Coverage, Low Income Subsidy enrollment, Retiree Drug Subsidy, and state-reported dual eligibility (Medicaid) status. As of March 2010, data formerly contained in Denominator Files were rolled into the Beneficiary Summary Files.
MedPAR	MedPAR files contain inpatient hospital and SNF final action stay records. Each MedPAR record represents a stay in an inpatient hospital or SNF, summarizing all services a beneficiary received from the time of admission through discharge; MedPAR records may represent one claim or multiple claims. The record unit of MedPAR files is the hospital or SNF stay.
Carrier Claim Files	Formerly called the Physician/Supplier Part B File, Carrier Claim Files contain final action claims data submitted by noninstitutional providers including physicians, physician assistants, clinical social workers, nurse practitioners, independent clinical laboratories, ambulance providers, and freestanding ambulatory surgical centers. Information contained in these files includes diagnosis and procedure codes (ICD-9-CM diagnosis, HCPCS), dates of service, reimbursement amount, noninstitutional provider numbers (e.g., UPINs, Provider Identification Numbers [PINs], National Provider Identifiers [NPIs]), and beneficiary demographic information. Each observation is at the claim level.
Hospice Claim Files	Hospice Claim Files contain final action claims data submitted by hospice providers, including the level of hospice care received (e.g., routine home care, inpatient respite care), terminal diagnosis (ICD-9-CM code), dates of service, reimbursement amount, hospice provider number, and beneficiary demographic information. Each observation is at the claim level.

EXHIBIT 5.2 (Continued)

Part D Drug Event (PDE) Files	Part D Drug Event Files contain prescription drug costs and payment data that CMS uses to make payments to the Part D plans and otherwise administer the Part D benefit. When a beneficiary fills a prescription under Medicare Part D, a prescription drug plan must submit a summary record to CMS. PDE data do not represent individual drug claim transactions but are summary extracts defined by CMS.
Non-Identifiable Data Files	
Part B Carrier Summary Data Files	Part B Carrier Summary Data Files are summarized at the carrier level by meaningful Healthcare Common Procedure Coding System/Current Procedural Terminology (HCPCS/CPT) code ranges. Each data set displays the allowed services, allowed charges, payment amounts by HCPCS/CPT codes, and prominent modifiers (2005–2008).
Provider of Service (POS)	Provider of Service (POS) extracts are generated from the Online Survey and Certification Reporting System (OSCAR) and include provider number, name, and address of participating institutional provider (1991 through current calendar quarter).
Medicare Cost Reports (MCRs)	Medicare-certified institutional providers must submit to a fiscal intermediary an annual cost report containing provider information, such as facility characteristics, utilization data, cost and charges by cost center (in total and for Medicare), Medicare settlement data, and financial statement data. CMS maintains cost report data in the Healthcare Provider Cost Reporting Information System (HCRIS). MCRs are available from 1996 through current calendar quarter.
HCRIS Reports	These reports are generated from HCRIS. They contain data on Medicare Part A (inpatient, SNF, home health agency, and hospice) and Part B (outpatient) institutional providers (2000–2008). Data in HCRIS are summarized from the SAFs.
UPIN Directory	The UPIN Directory contains selected information on physicians, doctors of osteopathy, limited-licensed practitioners, and some nonphysician practitioners who participate in Medicare (2003–2007). Data elements include UPIN, full name, specialty, physician license state code, zip code, Medicare provider billing number, and state.
Drug Plan Formulary	Prescription Drug Plan Formulary and Pharmacy Network Files contain formulary and pharmacy network data for Medicare prescription drug plans and Medicare Advantage prescription drug plans (with the exception of employer plans and the Program of All-Inclusive Care for the Elderly). These files are updated monthly. Updates are available at the end of the first complete week of each month (2005–2010).

*Table indicates years for which data are available.

Source: Data from ResDAC (2012).

However, states vary widely in the restrictiveness of their Medicaid eligibility requirements. The Personal Responsibility and Work Opportunity Act of 1996 "effectively decoupled Medicaid from cash assistance for low-income families" (Klemm 2000, 110). Medicaid will change significantly if ACA provisions take effect in 2014: Regardless of whether they were included in Medicaid's original entitlement categories, all non-elderly adults with incomes up to 133 percent of the federal poverty level (about $30,000 for a family of four) will become eligible for Medicaid (Ku et al. 2011). Whether this expansion of Medicaid will occur depends on the outcome of the US Supreme Court case *Florida v. Department of Health and Human Services* set to be argued in spring 2012.

Medicaid eligibility, benefits, and coverage standards vary across states, as do the databases. The 1997 Balanced Budget Act required all states to submit their Medicaid claims and enrollment records quarterly to the Medicaid Statistical Information System (MSIS) beginning in 1999. Although the MSIS contains a data dictionary defining required variables, states have been able to use their own coding schemes, especially for procedures. Medicaid administrative databases are now available for all 50 states. Starting with 1999 data, State Medicaid Research Files (SMRFs) have been replaced by Medicaid Analytic Extract files ("MAX" files). SMRFs and MAX files link enrollment information to claims, creating person-level records. Therefore, these databases are considered research identifiable, and researchers must get permission to obtain these data sets. The most efficient way for researchers to obtain these Medicaid data files is to go through ResDAC.

Medicaid data sets are typically available on a calendar year basis. These data files represent final, adjudicated claims, and the data sets have been extensively cleaned with a variety of edit checks (i.e., cleaning eliminates out-of-range or implausible values and other findings that suggest errors). MAX files contain one file with enrollment information and four additional claims files:

- Inpatient File: complete stay records for inpatient services, including up to ten diagnosis codes, up to seven procedure codes, discharge status, length of stay, and reimbursement amount
- Long Term Care File: claims for long-term care services provided by SNFs, intermediate care facilities, and independent psychiatric facilities (Variables include facility type, dates of service, and discharge status. This file contains up to five diagnosis codes but no procedure codes.)
- Other Therapy File: claims from all noninstitutional Medicaid providers, including physician services, laboratory and radiology services, clinic services, and premium payments (As appropriate, this file contains up to two diagnosis codes and one procedure code.)
- Drug File: final, adjudicated claims with an NDC code representing the drug (One procedure field is also present, but no diagnosis fields.)

Before the implementation of Medicare Part D, Medicaid data were the primary source of information about prescription drug use in large populations. Medicaid's computerized pharmacy data file contains information on all prescriptions dispensed by pharmacies, including the beneficiary's identification number, prescription date, drug (NDC code), quantity, prescribing physician, and reimbursement information. These data are reasonably reliable (Iezzoni 2002). One study found that 94 percent of pharmacy records matched claims in the Medicaid data set (Bright, Avorn, and Everitt 1989). Like Medicare Part D data, Medicaid pharmacy data represent prescriptions that were filled, not prescriptions that were written but never obtained.

Medicaid data represent the health care experiences of indigent persons, whose risk factors may differ from those of wealthier individuals (see Chapter 3). Risk adjustment methods derived for other populations may not apply equally well to Medicaid recipients. Nonetheless, detailed information on this vulnerable population is an important strength of Medicaid data. The ACA-mandated addition of many new Medicaid recipients will likely change the nature of this population. One challenge in conducting research with Medicaid populations is *churn*—the on again, off again eligibility for the program that causes recipients to appear and disappear periodically from the data sets. Churn could also affect the availability of data on newly added ACA-mandated Medicaid recipients as persons' incomes waver just above or below the 133 percent of the federal poverty level target. One study suggested that over the course of a year, Medicaid eligibility status could change for 50 percent of persons with incomes near the federal poverty level (Sommers and Rosenbaum 2011). Thus, churn in Medicaid populations could present significant difficulties for researchers analyzing Medicaid data.

Veterans Health Administration

The Department of Defense, Office of Civilian Health and Medical Program of the Uniformed Services (CHAMPUS), and affiliated TRICARE health plans maintain an automated information and reporting system primarily to manage their health care delivery system, which cares for active duty military personnel, National Guard and Reserve members, military retirees, and the families of all of these groups. CHAMPUS and TRICARE information is available on a limited special-request basis to support service or command projects, not for general health services research. In contrast, the VA maintains an active health services research program, drawing extensively from the computerized files created through administering Veterans Health Administration (VHA) hospitals and clinics.

The VIReC (VA Information Resource Center) website—known as the "Researcher's Guide to VA Data" (www.virecresearch.va.gov)—reminds visitors that the VHA is the "largest integrated health care delivery system in the U.S., comprising 162 VA hospitals, 137 nursing homes, 43 domiciliaries, and

more than 850 community- and facility-based clinics" nationwide (VIReC 2012a). VA facilities serve selected individuals among the United States' 23 million military veterans. To determine eligibility for VHA services, discharged service members must apply for health benefits. Veterans are assigned to one of eight priority groups on the basis of whether they have a service-connected disability, their health, their income, years of service, and other factors. Decisions about the cutoff for eligibility by priority group are made periodically by balancing available resources against demand for services.[6]

As described in Chapter 6, the VA has been on the forefront of implementation of electronic health records and system-wide health information technology infrastructure development. The VA also produces standard administrative data, which researchers have used for many years. As noted on the VIReC website, the VA has also merged data on its enrollees with their Medicare claims, enabling investigators to track individuals across the VA and private-sector health care system reimbursed by Medicare.

The VA compiles data in its enterprise-wide Corporate Data Warehouse (CDW), which serves as a repository for health information from the VHA's clinical and administrative information systems. The CDW, which stores information in a relational database, aims to provide data and the tools required to support management functions, performance measurement, and research needs. Because many VA data sources are merged, groups of patients can be identified by various attributes, such as their diagnosis or procedure codes from inpatient and outpatient records and even clinical variables (e.g., blood pressure, height and weight), and by time frame. The data are timely; for example, information from inpatient and outpatient encounters is updated daily, as are various physiological and laboratory values. Most impressive is the CDW domain called Vital Signs, drawn from the VA's electronic health information system. Exhibit 5.3 lists examples of the clinical information available in the Vital Signs domain, which is updated daily from

EXHIBIT 5.3 Examples of Variables Available from VA CDW Vital Signs

Temperature	Audiometry
Blood pressure, systolic and diastolic	Hearing
Respiratory rate	Height
Central venous pressure	Weight
Pulse oximetry	Pain
Ventilator minute volume	Circumference/girth
Ventilator tidal volume	Fundal height
Vision, uncorrected	Fetal heart rate
Vision, corrected	Tonometry

Source: Information from VIReC (2012b).

sources nationwide. Text documents from the VA radiology and microbiology systems are expected to be added to the CDW shortly.

The VHA Office of Information at the Austin Information Technology Center (the central repository for VA data) produces data sets in SAS for researchers to use. The data are drawn from the National Patient Care Database, which receives updated encounter data daily from VHA care sites nationwide and compiles them into a relational database. The Medical SAS Inpatient Datasets, sometimes called the Patient Treatment File, encompass four main categories of care: acute, extended, observation, and non-VA. Each of these categories comprises four databases: main (episode of inpatient care), bed section (care provided by specialty services), procedure (one day's procedures during an inpatient stay), and surgery (one day's surgeries during an inpatient stay). The Medical SAS Outpatient Datasets, sometimes called the Outpatient Care File, include two databases: the visit file, which represents up to 15 clinical stops in a given day, and the event file, which represents a single outpatient encounter, including date, appointment time, provider type, procedure codes (up to 15), and surgery codes (up to 15) (VA 2012).

The VHA has historically retained slightly different information than other public and private health care systems. For example, the VHA defines the first-listed hospital diagnosis differently than did UHDDS. Medicare and the UB-04 use UHDDS instructions that the "principal diagnosis is designated and defined as: the condition established after study to be chiefly responsible for occasioning the admission of the patient to the hospital for care." In contrast, the VHA uses a "primary diagnosis," the condition that was primarily responsible for the length of the hospitalization.[7] Although principal and primary diagnoses are generally identical, differences can hamper comparisons of diagnosis data between VHA and UB-04-based data systems. Furthermore, unlike most other health data systems, VHA data do not contain dollar claims for various services (their inclusion is unnecessary given global budgeting). Therefore, cost analyses based on VA data must include proxies. In an analysis comparing risk adjusters using VA data, Rosen and colleagues (2001a) used two proxies for costs: ambulatory provider encounters and service days (the number of ambulatory visit days plus the number of inpatient days) over one year.

VA data are available only to certain researchers; investigators must spend a specified percentage of their weekly job hours working at the VA to obtain them. For this reason, risk adjustment studies using VA data have been performed by a select handful of VA-based investigators. Nonetheless, VA data sets offer a rich opportunity for exploring risk adjustment methods for a population that historically has borne a heavy burden of chronic and comorbid illness. As described in Chapter 6, the addition of clinical information to administrative data creates a unique data source that can produce

special insights into the marginal contribution of clinical variables to risk adjustment methods.

Private Health Insurance

As part of processing claims and encounter records, private health insurers also create administrative databases. These files are structured to support the business needs of insurers, not outside investigators, and they can be complicated. Nonetheless, several companies have obtained files from private health insurers for analytic use, including for use by outside investigators. These companies sell these large, multiorganizational databases to external parties, attaching strictures about their use and confidentiality of identifiable information. The data sets may cost up to tens of thousands of dollars.

Private health insurance claims offer information on services used by employed adults and their children—populations not represented in public administrative databases. Insurance files typically include details about plan enrollment and each covered service; some include additional information, such as worker attendance and specifics about plan design.

Apart from the technical and logistical hurdles of using private insurance claims files for population-based studies, questions arise concerning the scope and content of these files. Claims files do not reflect utilization of services for which enrollees do not submit claims. Some plans require enrollees to pay high deductibles; patients may not submit claims until they meet their deductible limits. Certain insurers retain information only on paid claims, deleting data on services contributing toward the deductible or exceeding maximum benefit levels (ACA provisions have eliminated caps on health insurance benefits). Claims from managed care organizations may represent only services rendered by outside providers that submit claims because their own clinicians are salaried or have other payment arrangements. Information on mental health care or prescription drugs may be incomplete, especially if these services are carved out of the insurance plan or handled by another organization (e.g., a pharmacy benefits management company).

The content of private insurance claims reflects the mechanism for claims submission and payment. For instance, providers may submit bundled bills for a series of services that may have been provided on different dates. Dates retained in the claim file may represent the date of payment or payment adjustment, not the date of service. Depending on billing procedures, the provider number listed on the claim may represent an individual physician, a group of doctors, or an institution. Private insurers often require fewer diagnoses to be stated on claims than do public programs. In addition, few private insurers collect information on race or ethnicity of their enrollees. HIPAA provisions have standardized many aspects of the health insurance claims environment, improving the ability of researchers to use these files.

Coding Nomenclatures

When used for examining health care outcomes, administrative data must contain credible clinical information about risk factors (see Chapter 3). According to skeptics, administrative data fall short here. Beyond age and sex, most clinical information in administrative files comes from coded diagnoses. As described in the next section, the World Health Organization (WHO) governs worldwide efforts to code and classify health conditions, continually reviewing and revising its flagship nomenclature, the ICD. By international treaty, countries report mortality figures using the ICD; the United States has reported mortality causes using the tenth edition (ICD-10) since January 1, 1999 (Iezzoni 2010). For morbidity reporting, the United States still uses ICD-9-CM, which it adopted in 1979.

As I write this chapter, the more than 30-year reign of ICD-9-CM is nearing its end. On August 22, 2008, HHS published a proposed rule to replace ICD-9-CM with ICD-10-CM for diagnoses and to replace the ICD-9-CM procedure listing with ICD-10-PCS for electronic health information transactions covered by HIPAA. After a comment period, HHS published final guidelines on January 16, 2009, specifying an anticipated implementation date of October 1, 2013 (HHS 2009; Iezzoni 2010). (See discussion about ICD-10 at the end of this chapter.) The federal government has delayed this change multiple times because of concerns about the feasibility and costs of the move; as this book goes to press, CMS is proposing to push implementation compliance to 2014. The new classification has different organizational structures and thousands more codes than does ICD-9-CM. The change will require intensive staff and clinician training and software modifications. In addition, critical administrative code-based algorithms, such as the Medicare Severity-DRGs used for hospital payment and the Hierarchical Condition Categories (HCCs) used for Medicare Advantage plan payment, will need reprogramming and recalibration (Iezzoni 2010).

Skeptics question the plan to move to ICD-10-CM and ICD-10-PCS in 2013. According to John D. Halamka, MD (2010, 598), chief information officer at Beth Israel Deaconess Medical Center in Boston:

> . . . CMS is requiring all health care stakeholders to move from . . . ICD-9-CM, with its 17,000 different codes, to ICD-10-CM, with 155,000 different codes, in 2013. . . . Many organizations believe that this migration will be expensive and burdensome, and will have limited impact on clinical care. I . . . recommend that the Interim Final Rule include changes to the current ICD-10 "migration" plan, although this recommendation is controversial.

Although this uncertainty lingers, others are gearing up for this major transition, which will have considerable implications for risk adjustment methods that rely on ICD codes after 2013—or 2014, if CMS proceeds with its proposed delay.

Origins of ICD and ICD-9-CM

ICD traces its roots to the First Statistical Congress in Brussels in 1853 and a multinational agreement on the need for consistent coding of causes of death. Two years later in Paris, the congress adopted general disease classification principles proposed by William Farr (introduced in Chapter 1), grouping conditions primarily by anatomical site (Israel 1978). Following this basic schema, in 1893 the International Statistical Institute (the successor to the congress) produced the *Classification of Causes of Death*, urging revisions every ten years. This lexicon evolved into the ICD, maintained by WHO since the 1940s. WHO convened the International Conference for the Ninth Revision in Geneva in 1977, and ICD-9 emerged in January 1979.

Although the ninth revision of ICD contained more than three times as many codes as the eighth, American clinicians believed that ICD-9 provided insufficient detail for US administrative uses (Slee 1978). As a result, the United States took its own unique approach toward coding for health care delivery system applications. Professional and provider associations worked with staff from the National Center for Health Statistics (NCHS), which oversees ICD diagnosis coding in the United States, to "clinically modify" ICD-9 for broadly defined clinical purposes. The resultant ICD-9-CM initially contained more than 14,000 codes (additional codes have been inserted over the years). Adding a fifth digit to many diagnosis codes allowed more clinical detail.

In the United States, official coding guidelines are promulgated by the four Cooperating Parties: the Coordination-Maintenance Committee at NCHS, CMS, the American Hospital Association (AHA), and the American Health Information Management Association (AHIMA). This same partnership is in place for ICD-10-CM and ICD-10-PCS. NCHS retains primary responsibility for diagnosis codes, ensuring that all changes comply with WHO policies and reflect technical or scientific advances in medical knowledge. Procedure codes are used primarily by the United States; therefore, changes in procedure coding need not pass through WHO. CMS has authority for procedure coding revisions and updates, responding to technological advances and reimbursement concerns. AHA publishes the *Coding Clinic*, which explains official coding guidelines promulgated by the four Cooperating Parties. AHIMA plays a major role in addressing implementation issues raised by the coding schemes.

Organization and Format of ICD-9-CM

ICD-9-CM has three volumes: Volume I, Diseases, Tabular List; Volume II, Diseases, Alphabetic Index; and Volume III, Procedures, Tabular List and Alphabetic Index. The addition of procedures to ICD-9-CM is uniquely American; WHO does not maintain procedure nomenclatures. This discussion concentrates on diagnosis coding because procedures are less commonly used for risk adjustment. One obvious exception is the DRGs, which sort

surgical patients into groups by procedures. Medicare Hospital Compare mortality reports consider previous procedures (e.g., prior coronary artery bypass graft [CABG] surgery) in their risk adjustment methodology. (Hospital Compare is discussed in several later chapters in this book.) One risk adjustment method uses codes for durable medical equipment to capture risks related to disability (van Kleef and van Vliet 2010).

ICD-9-CM Volume I presents diagnosis codes in a tabular list organized into 17 broad categories. Some categories reflect the original grouping by anatomical locations, whereas others represent pathophysiological perspectives (e.g., infectious and parasitic diseases, neoplasms) and broad categories of conditions (e.g., "symptoms, signs, and ill-defined conditions"). ICD-9-CM codes include up to five digits. Three-, four-, and five-digit codes are used as necessary to represent increasing levels of specificity. For example, the code for pneumococcal pneumonia includes three digits (481), indicating that further detail is not needed. In contrast, the coding of diabetes mellitus spans five digits, representing end organ involvement and type of diabetes.

Two supplementary classifications represent various factors affecting health and other needs. In the first supplementary classification, codes up to five digits starting with the letter V portray factors influencing health status and contact with health services, such as personal history of penicillin allergy (V14.0), noncompliance with medical treatment (V15.81), outcomes of delivery (such as single liveborn, V27.0), and unemployment (V62.0). Other V codes indicate that services rather than diagnoses prompted the health care encounter, such as routine infant or child health check (V20.2), chemotherapy (V58.1), and examination for medicolegal reasons (blood alcohol tests, V70.4).

A second supplementary classification lists environmental events, circumstances, and conditions that have caused injury, poisoning, or other adverse events. The codes, which contain up to five digits and start with the letter E, aim to provide details about patients' conditions, but they cannot be listed as the reason for admission or principal diagnosis under current guidelines. E codes convey detailed descriptive information—for example, crew of commercial aircraft involved in accident at takeoff or landing (E840.2), accidental poisoning by benzodiazepine tranquilizer (E853.2), excessive cold due to weather conditions (E901.0), child abuse by parent (E967.0), and bite by centipede or venomous millipede (E905.4). When the Federal Bureau of Investigation (FBI) determines that a condition was caused by terrorism, a code from the category E979 should be the only code assigned; unless the FBI has confirmed the cause as terrorism, an E979 code cannot be assigned (NCHS 2010b, 85).

The portion of the tabular list of diseases devoted to symptoms, signs, and ill-defined conditions portrays a broad array of conditions, as well as abnormal results of various diagnostic tests. Current coding rules stipulate

that signs or symptoms linked definitively to a specific diagnosis generally should not be coded; codes for definitive diagnoses should supersede them. Signs and symptoms should be listed, however, when a definitive diagnosis is not established or when the symptom represents "important problems in medical care." Examples of these codes include alteration of consciousness (780.0), abnormality of gait (781.2), anorexia (783.0), precordial chest pain (786.51), elevated sedimentation rate (790.1), and abnormal electrocardiogram (794.31).

Another set of codes depicts complications resulting from medical care (codes 996–999). Most of these codes specify the nature of the problem but not its causality (e.g., "bad luck" versus negligence). Examples include mechanical complication due to a heart valve prosthesis (996.02), postoperative shock (998.0), accidental puncture or laceration during a procedure (998.2), postoperative infection (998.5), foreign body accidentally left in patient during a procedure (998.4), and ABO blood type incompatibility reaction (999.6).

ICD-9-CM thus contains codes for many conditions that are technically not diseases. Given this diversity, creative, clinically based compilation of ICD-9-CM codes produces a picture of risk factors and patients' clinical

EXHIBIT 5.4 Discharge Abstract for a Woman Admitted Emergently to a California Hospital

Patient Data	Hospital Data	Admission Data	Disposition Data
ID: XXXXX Age: 64 Sex: female Race: white Residence zip code: XXXXX	Hospital ID: XXXXX Zip code: XXXXX	Length of stay: 23 days Admission type: emergency Source: emergency room Total charges: $35,201 Payer: Medicare	Disposition: home health service DRG: 150 (peritoneal adhesiolysis with complication or comorbidity)

DIAGNOSIS AND PROCEDURE CODES AND DAYS FROM ADMISSION TO PROCEDURE

Diagnosis Codes	Procedure Codes	Days
5696 enterostomy malfunction	4642 pericolostomy hernia repair	0
56081 intestinal adhesion with obstruction	545 peritoneal adhesiolysis	0
311 depressive disorder NEC	9112 culture—peritoneum	0
9974 surgical complication—gastrointestinal tract	545 peritoneal adhesiolysis	6
9985 postoperative infection	9608 insert (naso-) intestinal tube	6
	9112 culture—peritoneum	6
	8964 pulmonary artery wedge monitor	7
	8962 central venous pressure monitoring	7
	8763 small bowel series	6
	8744 routine chest x-ray	6
	8819 abdominal x-ray NEC	6
	9604 insert endotracheal tube	0
	9392 mechanical respiratory assistance NEC	0
	9052 culture—blood	0
	9043 culture and sensitivity—lower respiratory tract	0
	8952 electrocardiogram	

status. Exhibit 5.4 shows discharge abstract information taken from a patient admitted to a California hospital (Iezzoni et al. 1994a), suggesting the following story: A 64-year-old woman with a history of bowel resection came to the emergency room and was admitted emergently for surgical repair of a malfunctioning enterostomy. On the first hospital day, she underwent repair of a pericolostomy hernia and lysis of peritoneal adhesions. She developed postoperative gastrointestinal complications, and on the sixth hospital day she again underwent surgery for lysis of adhesions. She developed a postoperative infection and became acutely unstable. Starting on the seventh hospital day, she required monitoring of her central venous pressure and pulmonary artery wedge pressures. She finally needed endotracheal intubation and mechanical ventilation. By the twenty-third hospital day, she had recovered sufficiently to be discharged with instructions for home health care.

The detail and large numbers of ICD-9-CM codes can make them difficult to use in analyses. To make coded information easier to use, HCUP investigators at AHRQ (Elixhauser, Andrews, and Fox 1993; Elixhauser and McCarthy 1996) created the Clinical Classifications Software (CCS; www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp) by grouping numeric diagnosis codes, V codes, and procedure codes and eliminating codes that were exceedingly general (e.g., V15.81, noncompliance with medical treatment; 780.7, malaise and fatigue). CCS now comes in the following versions (readers should check for updated information, as this software is periodically revised):

- CCS for ICD-9-CM collapses the more than 14,000 diagnosis codes into 285 mutually exclusive, single-level groups; a multilevel version collapses these groups into hierarchies within broader body systems or condition categories.[8]
- CCS for Services and Procedures groups the roughly 3,900 procedure and service codes into 231 mutually exclusive, single-level groups; a multilevel version collapses these groups into hierarchies within broader body systems or condition categories.
- Since ICD-10 supplanted the ninth ICD revision for mortality reporting in 1999, in 2003 AHRQ released ICD-10 CCS for use with mortality data (www.ahrq.gov/data/hcup/icd10usrgd.htm). ICD-10 CCS has 259 mutually exclusive categories aimed at clinical homogeneity. Because of small numbers of cases, some heterogeneous conditions had to be grouped together to bolster sample sizes.

As described in Chapter 3, CCS is the basis for an increasingly popular approach to capturing comorbidity using administrative data (see Exhibit 3.6).

Coding Concerns

As noted at the outset of this chapter, administrative data are produced during health care delivery. Therefore, codes associated with administrative data reflect this provenance. These associations have raised concerns for decades among researchers using coded data. These concerns become especially pressing when using administrative data for risk adjustment related to payment or quality measurement purposes (Smith 2010, 126):

> In general, administrative databases contain prospectively collected demographic and financial information and retrospectively collected diagnostic and therapeutic information developed by professional hospital coders from chart review. . . . There are financial incentives to "upcode" diagnoses, a process that will degrade their utility in risk adjustment. The coding process is designed to detect all possible diagnoses, which are then frequently sorted by financial importance and truncated in transmission, further reducing the applicability of the risk profile created for the encounter. Because the coding process is not specific to an encounter or disease, the absence of a diagnosis does not definitively mean that it was not present, a distinct liability. Finally, many diagnoses used for risk adjustment are "synthesized" from financial events occurring in the encounter, through questionable methodologies.

Neither the Tabular List (Volume I) nor the Alphabetic Index (Volume II) offers explicit clinical definitions of ICD-9-CM codes. For example, ICD-9-CM includes about 40 four- and five-digit codes for different types of anemia, such as iron deficiency anemia secondary to inadequate dietary iron intake (280.1), thalassemia (282.4), autoimmune hemolytic anemia (283.0), and constitutional aplastic anemia (284.0). Nowhere, however, does it specify what level of hemoglobin or hematocrit justifies an anemia diagnosis. Similarly, the respiratory failure code (518.81) does not indicate what level of arterial oxygenation, respiratory rate or pattern, or other clinical abnormality merits that diagnosis. Coders assign ICD-9-CM diagnoses on the basis of physicians' documentation in medical records. Therefore, coding inevitably reflects the imprecision, inconsistencies, inaccuracies, and incompleteness of physicians' documentation patterns and terminology. Adding financial incentives to the mix complicates matters further.

The context of diagnosis coding in the United States changed with the 1983 enactment of Medicare's prospective payment system based on DRGs. In 1981, a prescient warning appeared in the *New England Journal of Medicine* (Simborg 1981, 1602, 1604):

> This article is intended to provide a case report of "DRG creep," a new phenomenon that is expected to occur in epidemic proportions in the 1980s. DRG creep may be defined as a deliberate and systematic shift in a hospital's reported case mix in order to improve reimbursement. . . . Minor diagnostic nuances and slight imprecisions of wording have little practical clinical importance, yet under DRG reimbursement they

> would have major financial consequences. . . . It is hoped that hospitals will refrain from disseminating the more virulent forms of DRG creep; however, the potential for a broad spectrum of manifestations certainly exists.

DRGs suddenly vested ICD-9-CM with a power for which it was never designed: determining hospital reimbursement. Under DRG-based payment, ICD-9-CM codes translate directly into dollars. As of April 1, 1989, Medicare also requires physicians to list ICD-9-CM diagnosis codes on their claims. New words, such as *optimization* and *maximization*, entered the coding vocabulary. Individual and institutional providers began focusing on coding practices, hoping to maximize their reimbursement. Some hospitals reorganized, moving medical record departments from general administration to financial divisions. Numerous companies sprung up to capitalize on this opportunity, producing software and providing coding support services.

Researchers and analysts using administrative data often assume that coding is easy and straightforward: Take the codebook off the shelf, search the index for the disease or condition, and voilà. That impression is incorrect. Coding, like the medical concepts and practice it aims to represent, is a science and an art. To become a certified coder, one must complete years of training and pass a certification examination. The Commission on Certification for Health Informatics and Information Management (CCHIIM) establishes and enforces standards and procedures for certification and recertification of health informatics and information management professionals. Candidates who graduate from a CCHIIM-recognized program are eligible to sit for the associated certification examination sponsored by AHIMA (CCHIIM 2011). AHIMA offers examinations for the following credentials:

- Certified Coding Associate (CCA)
- Certified Coding Specialist (CCS)
- Certified Coding Specialist—Physician-based (CCS-P)
- Registered Health Information Administrator (RHIA)
- Registered Health Information Technician (RHIT)
- Certified Health Data Analyst (CHDA)
- Certified in Healthcare Privacy and Security (CHPS)

Candidates for each examination must meet specific eligibility requirements. A number of other organizations also certify coders, such as the American Academy of Professional Coders, the Board of Medical Specialty Coding, and the Professional Association of Healthcare Coding Specialists.

In recent years, the field of health information management has grown tremendously, spurred by the expansion of the health care delivery system, efforts to manage costs, and oversight initiatives. A 2001 report by the US Bureau of Labor Statistics projected a growth of 49 percent in the number

of health information management and medical records jobs by 2011 (Hecker 2001, 68). However, the majority of these new professionals hold only associate's degrees and fall into the third quartile of median annual earning (i.e., low income). The impending move to ICD-10-CM is likely to open even more positions for coding professionals.

Overview of Coding Rules

ICD-9-CM general coding guidelines begin by emphasizing the partnership between coders and providers: "[A] joint effort between the healthcare provider and the coder is essential to achieve complete and accurate documentation, code assignment, and reporting of diagnoses and procedures. The importance of consistent, complete documentation in the medical record cannot be overemphasized" (NCHS 2010b, 95). In this context, *provider* means a physician or any qualified health care practitioner legally responsible for establishing patients' diagnoses.

Coding rules for inpatient and outpatient diagnoses differ. Coding hospital discharge diagnoses is typically more time consuming and complicated than coding diagnoses on outpatient claims. The two major steps in coding discharge diagnoses are (1) specifying the pertinent diagnosis codes and (2) determining their order (i.e., sequencing the codes). With few exceptions, the principal (first-listed) diagnosis drives DRG assignment and thus hospital reimbursement, underscoring the importance of the sequencing step. "Other" or secondary diagnoses are diagnoses that require clinical evaluation, therapeutic intervention or treatment, or diagnostic procedures; extend length of hospital stay; or increase nursing care or monitoring (NCHS 2010b, 89).

To decide which diagnoses to code, coders must review medical records in their entirety to identify diagnoses established or entertained by the physicians treating the patient. UHDDS guidelines cover coding admissions to short-term acute care hospitals and other settings of care (including long-term care facilities, psychiatric hospitals, home health agencies, rehabilitation facilities, and nursing homes). These guidelines demand coding of (NCHS 2010b, 89)

> . . . all conditions that coexist at the time of admission, that develop subsequently, or that affect the treatment received and/or the length of stay. Diagnoses that relate to an earlier episode which have no bearing on the current hospital stay are to be excluded. . . . Some providers include in the diagnostic statement resolved conditions or diagnoses and status-post procedures from previous admission that have no bearing on the current stay. Such conditions are not to be reported and are coded only if required by hospital policy. . . . Abnormal findings (laboratory, x-ray, pathologic, and other diagnostic results) are not coded and reported unless the provider indicates their clinical significance . . . .

UHDDS guidelines also address coding of uncertain diagnoses. The rules apply only to inpatient admissions of short-term acute care hospitals, long-term care facilities, and psychiatric hospitals (NCHS 2010b, 90):

> If the diagnosis documented at the time of discharge is qualified as "probable", "suspected", "likely", "questionable", "possible", or "still to be ruled out" or other similar terms indicating uncertainty, code the condition as if it existed or was established. The bases for these guidelines are the diagnostic workup, arrangements for further workup or observation, and initial therapeutic approach that correspond most closely with the established diagnosis.

Outpatient coding guidelines differ significantly from the UHDDS inpatient guidelines. First, the term *first-listed diagnosis* is used instead of *principal diagnosis*.[9] Second, whereas probable or possible diagnoses are permitted in inpatient coding, only confirmed diagnoses are permitted in outpatient coding (guidelines recognize that more than one visit may be required to confirm diagnoses) (NCHS 2010b, 92):

> Do not code diagnoses documented as "probable", "suspected," "questionable," "rule out," or "working diagnosis" or other similar terms indicating uncertainty. Rather, code the condition(s) to the highest degree of certainty for that encounter/visit, such as symptoms, signs, abnormal test results, or other reason for the visit. . . . This differs from the coding practices used by short-term, acute care, long-term care and psychiatric hospitals.

Thus, if no diagnosis has been clinically confirmed, codes for symptoms or signs are permitted in outpatient records.

Coding guidelines for outpatient encounters further specify the coding of all documented conditions that exist at the time of the encounter or visit and that require or affect treatment or management. Conditions that were previously treated and are no longer active cannot be coded, although history codes (V codes) can be coded as secondary diagnoses if the historical condition or family history affects current care. For outpatient visits involving the provision of diagnostic services only, coders are instructed to list first the diagnosis, condition, or problem "chiefly responsible for the outpatient services provided during the encounter/visit. Codes for other diagnoses (e.g., chronic conditions) may be sequenced as additional diagnoses" (NCHS 2010b, 93).

Present on Admission Indicators

One recent coding rule—the 2008 change requiring present on admission (POA) indicators for hospital discharge diagnosis codes (i.e., flags indicating whether secondary diagnoses were present on admission or arose subsequently during the hospital stay)—has particular implications for risk adjustment. POA flags enable risk adjustment methods to include or exclude

conditions that newly arise during the hospital stay, either as complications of patients' diseases or, more critically, complications of care. Depending on the purpose of the risk adjustment, this distinction may be conceptually critical. For example, risk adjustment for examining quality outcomes would want to exclude conditions occurring newly in hospital because they might reflect quality shortfalls (see Chapter 4).

The POA idea was considered for many years. In June 1992, NCVHS proposed multiple changes to UHDDS, including adding an "alpha" POA qualifier to each discharge diagnosis. NCVHS based this proposal on experiences at Mayo Clinic and in New York State, where POA indicators added "modest additional cost" to data collection. A dozen years later, NCVHS again recommended adding a discharge diagnosis modifier for POA, but the government never adopted NCVHS's proposals (Iezzoni 2007).

Medicare policy mandates then made adding POA indicators suddenly urgent (Iezzoni 2007). Section 5001(c) of the Deficit Reduction Act of 2005 required Medicare to stop paying hospitals for the costs of treating certain conditions that occur in hospital for discharges on or after October 1, 2008.[10] The goal was to eliminate Medicare payments to hospitals for complications that should have been prevented. POA indicators offered the only feasible option to support this new mandate (Coffey, Milenkovic, and Andrews 2006). Effective March 1, 2007, UB-04 added fields for POA indicators for principal and secondary discharge diagnoses following standard coding guidelines (Exhibit 5.5). By definition, principal diagnoses generally are present on admission because they represent the condition that caused the hospitalization. Coding rules for certain ICD-9-CM codes, such as V codes for chemotherapy, require that they be listed as the principal diagnosis; the POA concept clearly does not relate to such codes. Because of these few exceptions, UB-04 has POA slots for principal as well as secondary diagnoses. Coding guidelines also exempt some codes from requiring POA indicators, such as V codes for personal history of malignant neoplasm.

How much this new policy to eliminate payments for certain complications would reduce Medicare expenditures was unclear (Rosenthal 2007). Using hospital claims from California and New York, which had included POA indicators for over a decade, Zhan and colleagues (2007) investigated

EXHIBIT 5.5
Definitions of Present on Admission Indicators for Principal and Secondary Diagnoses

Y	Present at the time of inpatient admission
N	Not present at the time of inpatient admission
U	Documentation is insufficient to determine if condition is present on admission
W	Provider is unable to clinically determine whether condition was present on admission or not

this question by assigning DRGs using all diagnoses and then again using only POA diagnoses. Relatively few cases (1.4 percent) shifted to lower-cost DRGs using only POA diagnoses for DRG assignment, and projected nationwide cost savings were relatively modest (0.6 percent) compared with total Medicare hospital expenditures.

These small effects were not surprising. Today's hospitalized Medicare beneficiaries, with some exceptions (e.g., scheduled-surgery patients), are typically very sick on admission, warranting multiple secondary diagnoses flagged POA. For disabled and elderly Medicare beneficiaries, many other secondary diagnosis codes that drive assignment to higher-cost DRGs likely represent chronic conditions, which are obviously POA. In addition, findings for individual hospitals (e.g., percentage of POA diagnoses) suggested significant practical challenges to ensuring accurate assignment of POA indicators and raised serious concerns about coding consistency among hospitals (Zhan et al. 2007; Iezzoni 2007). Studies in California and New York also questioned the accuracy of these indicators (Coffey, Milenkovic, and Andrews 2006).

Accuracy and Reliability of Coded Data

As early as the 1970s, studies questioned the accuracy of diagnostic coding (IOM 1977a, 1977b). A review of nearly 2,000 hospital discharge abstracts found nearly 100 percent accuracy of basic information, such as admission and discharge dates, patient age and sex, and payer. However, the researchers and hospitals agreed on only 65.2 percent of the principal diagnosis codes (IOM 1977a). Many problems were related to the vagaries of ICD-8 (the ICD of that time) and its diagnosis codes, coding guidelines, and sequencing rules. For 10.7 percent of cases reviewed, discrepancies between the researchers' and hospitals' principal diagnoses were irreconcilable because of legitimate professional disagreements in interpreting the medical record and pertinent coding guidelines, particularly sequencing (76.5 percent of indeterminate cases). The investigators concluded, "One must assume that abstracted hospital data contain errors and use them with caution. The seriousness of the error depends on the purpose to which the data are applied" (IOM 1977a, 49). Nevertheless, within a few years, Medicare made diagnosis codes through DRGs the basis of hospital payment.

Implications of Linking Coding and Payment

One barometer reflecting the effect of payment on coding practices is the Medicare case-mix index (CMI). The CMI reflects the average relative weight of DRGs assigned to hospitalized patients. Watchdog groups and Medicare annually track the CMI for individual hospitals, looking for targets

to investigate. Higher CMIs represent higher aggregate hospital reimbursements caused by a shifting DRG mix of patients.

The CMI rose 32.4 percent cumulatively from the first year of prospective payment through 1996 (Prospective Payment Assessment Commission 1996, 61). The largest increases occurred in the two years after prospective payment was instituted (fiscal years 1984 and 1985): Expenditures per hospital discharge rose 18.5 percent and 10.5 percent, respectively, and payments increased by 5 percent to 6 percent annually up to the ninth year. These higher CMIs suggested that Simborg's 1981 prophecy of DRG creep had come true, but the exact causes of CMI changes proved difficult to untangle. The central question was whether coding changes represented creep or "optimization." Creep implied willful disregard of coding rules, whereas optimization suggested taking lawful advantage of coding opportunities and thus DRG assignment. Although code creep caused some portion of the CMI growth, other influential factors included changes in practice patterns (e.g., shifting less-ill inpatients to outpatient care, substituting surgical for medical services, developing and diffusing new technologies), heightened oversight of admissions to acute care facilities, aging of the population, hospital structural changes, and changes in the Medicare hospital payment program.

Soon after Medicare instituted DRGs for hospital payment, several studies attempted to disentangle the roots of CMI increases. The first study, conducted by RAND, found that two-thirds of the 2.4 percent increase in the CMI between 1986 and 1987 was real (i.e., patients really were sicker); the other third resulted from coding practice changes and modifications to the DRG classification system (Carter, Newhouse, and Relles 1990). A subsequent study found that half of the 3.3 percent increase in the CMI between 1987 and 1988 was real (Carter, Newhouse, and Relles 1991). Goldfarb and Coffey (1992) dissected real from coding changes in the CMI by looking at shifts in practice patterns and aging of the population. They found that hospitals had largely completed efforts to "improve" their coding practices by 1986.

In the 1980s, the HHS Office of Inspector General (OIG) conducted major studies of coding accuracy explicitly focused on coding for DRG assignment. The first OIG study sampled Medicare admissions to 239 hospitals nationwide between October 1, 1984, and March 31, 1985 (Hsia et al. 1988). An OIG contractor reviewed photocopied medical records and compared DRGs originally assigned by the hospitals to those assigned on the basis of reabstracted data. Of 7,050 cases, the OIG reabstraction changed DRG assignment for 1,374, an average of 20.8 percent (after reweighting results by hospital size). More important, 61.7 percent of discrepancies financially favored the hospitals, regardless of size. Given errors in DRG assignment, Medicare overpaid hospitals by an extrapolated $300 million in fiscal year 1985.

The OIG repeated its DRG validation study using 2,451 patient records from admissions of Medicare beneficiaries in 1988 (Hsia et al. 1992).

In this sample, 14.7 percent of records contained errors that changed DRG assignment, but this time only 50.7 percent of the coding errors financially benefited the hospitals. Misspecification remained the most common cause of errors (62.9 percent), but the proportion of misspecifications that financially favored the hospitals decreased significantly from the 1985 level. However, incorrect sequencing continued to contribute to excessive hospital payment.

Anecdotal reports suggested that as hospitals faced financial pressures, they tried to enlist physicians' assistance in optimizing coding. Some hospitals posted notices in physicians' lounges documenting the financial ramifications of coding. One hospital distributed an e-mail message to its medical staff urging them to use specific terminology rather than vague phrases that cannot sustain codes (e.g., to write "blood loss anemia" rather than "low hematocrit"). To prevent DRG creep, from the mid-1980s until September 1995, federal regulations required that physicians "attest" in writing to the accuracy of the discharge diagnoses of Medicare patients before hospitals submitted bills for payment. In 1995, however, Vice President Al Gore eliminated the attestation requirement as part of his broader "reinventing government" initiative, which aimed to simplify federal paperwork. Nonetheless, within a year, worries resurfaced about potential upcoding. Federal regulators found that hospitals hired consultants explicitly to maximize reimbursement through "enhanced" coding. In 1996 HIPAA stipulated strict civil penalties for upcoding.

Medicare is taking a more aggressive stance toward the implications of what is now called—perhaps euphemistically—"increased coding intensity" involving managed care plans. According to the Medicare Payment Advisory Commission (MedPAC 2011a, 291), enrollment in Medicare Advantage (MA) plans in November 2010 was 11.4 million beneficiaries, or 24 percent of all Medicare beneficiaries. Medicare calculates capitated payments to MA plans for each enrollee separately on the basis of that individual's diagnoses in the previous year. FFS beneficiaries' diagnostic information comes from FFS claims; MA enrollees' diagnostic information comes from their MA plans. These diagnoses are used to produce HCC risk scores, which are then used to set payment levels. However, diagnoses for MA plan members grew more rapidly than did those for FFS beneficiaries. To take into account several years of these presumably excessive coding increases, CMS reduced the risk scores by 3.41 percent for 2010 and 2011 (MedPAC 2011a, 309).[11]

Other Studies of Coding Accuracy

Most studies of diagnosis coding accuracy have been performed in the context of payment policy, for obvious reasons. But other smaller studies have explored specific topics, most notably issues relating to using coded data to identify quality problems. Given the imminent shift to ICD-10-CM and the

likelihood that change will fundamentally alter the coding landscape, I do not review current or prior coding accuracy studies in depth here. Instead I highlight several key areas relevant to risk adjustment.

Among the states, California and its Office of Statewide Health Planning and Development (OSHPD) have been at the forefront of monitoring hospital coding for many years (Meux, Stith, and Zach 1990). In one study, OSHPD reabstracted 2,579 records from discharges from 30 randomly selected hospitals in 1988 across a variety of conditions (e.g., obstetrics, newborn, and psychiatric conditions). The study found that coding accuracy varied by condition, ranging from 54.5 percent for old myocardial infarction to 97.9 percent for pneumonia and acute cerebrovascular disease (Romano and Luft 1992). Another California investigation involving reabstraction of 974 inpatients coded as having acute myocardial infarctions (AMIs) found problems even with the coding of important secondary diagnoses. Coding quality was particularly poor for hypotension, pulmonary edema, other valve disease, nutritional deficiency, chronic liver disease, and late effects of cerebrovascular disease (Wilson, Smoley, and Werdegar 1996).[12] California stopped adjusting for these poorly coded conditions in its report card on hospitals' AMI mortality rates.

This reabstraction study of 974 AMI and another 991 diskectomy patients admitted to California hospitals in 1990 and 1991 was prompted by the state's initiative to produce hospital report cards on AMI mortality and diskectomy complication rates using administrative data (Wilson, Smoley, and Werdegar 1996). Not surprisingly, these reports attracted criticism because of their administrative data source. For the study to validate coding accuracy, California analysts sampled cases from hospitals stratified into three groups: hospitals whose risk-adjusted AMI mortality was (1) better than expected, (2) neither better nor worse than expected, and (3) worse than expected. Overall, 65.0 percent of discharge abstracts for AMI cases were missing at least one clinical risk factor; 30.9 percent were missing two risk factors. Hospitals varied from 45 percent to 87 percent in the fraction of uncoded risk factors, but the percentage of missing risk factors did not vary across the three hospital mortality categories. In contrast, overcoding rates varied widely across the three categories. Of the discharge abstracts, 31.5 percent contained at least one "unsupported risk factor." This overcoding was much more common at low-mortality hospitals than at intermediate- or high-mortality hospitals (36.7 percent versus 29.2 percent and 29.0 percent, respectively; $p = 0.04$). Overcoding rates ranged from 10 percent at one high-mortality hospital to 74 percent at a low-mortality hospital. Variation in coding accuracy explained part of the differences between high- and low-mortality hospitals. Similarly, underreporting of diskectomy complications was much higher at hospitals with lower-than-expected complication rates (Romano et al. 2002).

Lawthers and colleagues (2000) examined secondary diagnosis (and some procedure) coding by reabstracting more than 1,200 medical records

from California and Connecticut, but their reabstractions had a twist. After trained coding technicians completed their reviews using specially designed software, they viewed the codes originally assigned by the hospital and decided whether they wished to add other codes. The codes of about 85 percent of the cases were confirmed during the initial (blinded) review, and the coding technicians added new codes to the other 15 percent after seeing the hospitals' original codes. Interrater reliability among the coders was only moderate. Some codes (e.g., for AMI, stroke, hemorrhage, pulmonary embolism) were more likely than others to be identified during the initial review and to demonstrate excellent interrater reliability.

Researchers have used longitudinal data to examine the completeness of coding of chronic, incurable conditions. Of Medicare beneficiaries coded with an inpatient or outpatient diagnosis of dementia in 1994, only 59 percent had dementia coded in 1995; of patients coded with paraplegia or quadriplegia in 1994, only 52 percent had these codes in 1995 (MedPAC 1998, 17). Using Medicaid data from seven states, Kronick and colleagues (2000, 60) found that the following percentages of people with a specific diagnosis coded in a given year had the same diagnosis coded in the following year: 80 percent for schizophrenia, 68 percent for diabetes, 58 percent for multiple sclerosis, 57 percent for quadriplegia, and 34 percent for cystic fibrosis. As these conditions do not disappear, the absence of their codes on patients' claims in the subsequent year underscores the incompleteness of coding (Iezzoni 2010). When analysts use longitudinal files, they can identify comorbid conditions during some period before hospitalization (e.g., pneumonia present several days prior to hospitalization), but they must recognize that the rules for coding outpatient diagnoses differ from the rules for inpatient coding (see discussion earlier in the chapter). These differences between inpatient and outpatient coding rules might explain why some presumably chronic conditions coded in one setting are not also coded in the other.

Some conditions are rarely reported, even if they exist. Coding guidelines stipulate that unless conditions are actively addressed, they should not be coded. Although codes exist for blindness, deafness, and hard of hearing, for example, hospitals and physicians infrequently list these conditions on claims for services unrelated to the eyes or ears. Mental retardation is rarely coded for adults or children, perhaps because few health interventions directly treat this condition (Perrin et al. 1999; Kronick et al. 2000). Sometimes physicians intentionally withhold codes for potentially stigmatizing diagnoses (e.g., mental health disorders) when they can legitimately list other conditions.

Little information exists about the quality of coded diagnosis data in other inpatient facilities (e.g., psychiatric or rehabilitation hospitals) or outpatient settings. A comparison of Medicare Part B claims and physicians' office records for 1,596 Maryland patients from 1990 to 1991 produced worrisome results. The Part B claims and office records matched on only

40.3 percent of zip codes and 58.5 percent of birth dates (Fowles et al. 1995, 192). The kappa statistics (see Chapter 9) indicating the level of agreement ranged from 0.0 (ketoacidosis) to 0.72 (diabetes mellitus) for the 26 diagnoses examined and were greater than 0.40 for only six diagnoses. Many claims did not indicate diagnoses noted in the record.

A study published by researchers from the Dartmouth Atlas team (Welch et al. 2011) sought to determine the association between frequency of diagnoses for chronic conditions in geographic areas and case-fatality rate among Medicare beneficiaries. Although it was not a study of coding accuracy, its provocative findings have implications for risk adjustment:

- Across 306 hospital referral regions, the frequency of diagnosis codes varied widely for Medicare FFS beneficiaries, from 0.58 chronic conditions in Grand Junction, Colorado, to 1.23 in Miami, Florida.
- As regional diagnosis frequency increased, the case fatality associated with a chronic condition decreased; thus, an inverse relationship exists between the regional frequency of chronic condition diagnoses and their case-fatality rate.

What might explain these seemingly paradoxical findings? Systematic variations in coding practices by region, potentially motivated by "gaming" (i.e., efforts to increase reimbursement), could explain these results; however, no evidence exists to support this possibility. Another possibility is that regions with high diagnosis frequency are more effective at treating chronic conditions and thus have lower condition-specific mortality. Yet another explanation is that frequency differences relate to differences in the intensity of physician encounters and diagnostic testing by region: "More testing and more opportunities to make diagnoses may translate into the typical patient given a diagnosis being less sick" (Welch et al. 2011, 1117). In other words, in regions with high rates of testing and more frequent physician visits, an average person is more likely to have one or more diagnoses coded from these numerous encounters; in contrast, in regions with less frequent physician contacts, only sicker-than-average individuals receive services and thus have diagnoses coded. On the basis of their findings, the researchers urged caution in using diagnoses from claims data for risk adjustment (Welch et al. 2011, 1118):

> If diagnosis is not solely an attribute of underlying disease burden, adjustments based on frequency of diagnosis may introduce bias into efforts to compare outcomes, pay for health care, and assess the extent of geographic variation in health care delivery. On the other hand, if more diagnoses (and more frequent encounters and diagnostic testing as well as greater spending) improve outcomes, then standard methods of risk adjustment may provide a more accurate comparison of effectiveness and efficiency. Future research must further evaluate the contribution of the process of observation to diagnosis frequency and explore mechanisms to better measure disease burden.

Other Implications of Using Coded Data for Risk Adjustment

Beyond concerns about their completeness, accuracy, and clinical meaningfulness, coded data present other challenges for risk adjustment. As noted in the POA discussion above, questions about timing of diagnoses, especially for hospitalized patients, raise important concerns. In our work on hospital-based severity measures as shown in Exhibit 5.6, discharge abstract–based risk adjusters were sometimes equal or better statistical predictors of in-hospital mortality than were adjusters derived from more clinically meaningful data abstracted from medical records, notably clinical indicators on admission to hospital (Iezzoni 1997; Iezzoni et al. 1995a, 1995b, 1996a, 1996b; Landon et al. 1996). When examining CABG mortality in New York State using MedPAR versus clinical data, Hannan and colleagues (1997) found similar results: c-statistics for models fell when they eliminated ICD-9-CM codes representing complications rather than comorbidities. One hypothesis about why discharge abstract–based risk adjusters perform well in predicting in-hospital deaths is their reliance on ICD-9-CM codes for conditions, such as cardiac arrest, arising late in the hospital stay (Iezzoni et al. 1996d). In contrast, the clinically based measures use findings from only the first two hospital days. One conjecture is that discharge abstract–based measures benefited from a virtual tautology, using near-death experiences to predict death. The POA flag could rectify this problem.

Differentiation of POA from non-POA diagnoses may significantly affect perceptions of relative hospital mortality rates. Glance and colleagues (2008) analyzed data from 2.1 million California hospital discharges between 1998 and 2000, calculating risk-adjusted mortality rates using either all discharge diagnoses or only diagnoses with POA flags. When they used only

	Condition			
Severity/Measure	**AMI**	**CABG**	**Pneumonia**	**Stroke**
Clinically Based Measures				
MedisGroups empirical version	0.83	0.74	0.85	0.87
Physiology score 2	0.83	0.73	0.82	0.84
Discharge Abstract–Based Measures				
APR-DRGs	0.84	0.83	0.78	0.77
DS mortality probability	0.86	0.78	0.80	0.74
Patient Management Category severity score	0.82	0.81	0.79	0.73

Source: Adapted from Iezzoni et al. (1995b, 1996b, 1996c); Landon et al. (1996).

EXHIBIT 5.6 c-Statistics for Predicting In-Hospital Death

diagnoses with POA indicators, the relative ranking of hospitals as high or low mortality frequently changed: 27 percent (stroke) to 94 percent (CABG) of hospitals classified as high performing on the basis of all diagnoses for risk adjustment were classified as either intermediate or low performers when only POA diagnoses were used, and 25 percent (heart failure) to 76 percent (percutaneous coronary intervention) of hospitals classified as low performing on the basis of POA diagnoses were classified as intermediate performers when all diagnoses for risk adjustment were used. These findings raise serious questions about risk adjustment methods that fail to account for diagnosis timing when using mortality performance to draw inferences about hospital quality.

Despite the value of POA indicators, questions remain about their accuracy. Pine and colleagues (2009a) examined the plausibility of POA coding in various clinical contexts (different types of hospitalizations) using three years of discharge abstract data from New York State, which implemented POA coding in 1993. The investigators found that POA flags were frequently problematic. Furthermore, variation in POA coding practices across hospitals was substantial. This interhospital variability could compromise comparisons of risk-adjusted outcomes based on POA diagnoses.

Another concern about using administrative data for risk adjustment is that key clinical information may be missing. As noted earlier, coded data may not provide sufficient detail about clinical complexity or acuity of patients' conditions. Also unclear from discharge abstracts are the goals of patients' hospitalizations. Goals matter enormously when predicting some hospital outcomes, most notably imminent deaths. According to Holloway and Quill (2007, 802), "Mortality has been criticized as a measure of quality for years and debates about methods of risk adjustment are almost clichéd," but neglected in these debates are concerns about "preference-sensitive care." Hospitals vary widely in the use of early do not resuscitate (DNR) orders, and hospital mortality measures erroneously treat all deaths as medical failures. For example, Hospital Compare (www.hospitalcompare.hhs.gov) identified a Buffalo, New York, hospital as one of the 35 worst US hospitals because its July 2005 to June 2006 congestive heart failure (CHF) mortality rate was 4.9 percentage points higher than the national mean. The hospital reviewed medical records of its CHF deaths and found that 11 decedents (about 40 percent of total CHF mortalities) were in hospice or receiving palliative care–only treatment at the patients' requests (Holloway and Quill 2007, 802).

An ICD-9-CM code (V66.7) exists for palliative care (Cassel and Vladeck 1996), and some discharge abstracts include information on hospice services and DNR status. Nonetheless, this information is not used consistently in risk-adjusting hospital mortality rates. One study examined how the use of hospice and palliative care is considered by prominent producers of hospital mortality rates for public reports: CMS Hospital Compare, *U.S.*

News & World Report "Best Hospitals," Thomson Reuters "100 Top Hospitals," and HealthGrades (Cassel et al. 2010). All methods used standard discharge abstract data, but they recognized and addressed cases involving hospice or palliative care differently. One method eliminated cases with prior hospice care; another excluded patients discharged to hospice after their index hospitalization; and two eliminated some or all cases with the V66.7 palliative care encounter diagnosis code.

Merging Administrative Data and Other Information Resources

As described further in Chapter 6, one way to enhance administrative data is to link them with other data sources—for example, merging administrative data with clinical information from medical records. Pine and collaborators (2009b) used claims and data abstracted from medical records from 188 Pennsylvania hospitals to examine risk adjustment for five medical conditions and three surgical procedures. The investigators produced a series of risk adjustment models, each of which used the following information differently: POA indicators; additional diagnoses abstracted from the medical record but not coded on the claim; numerical laboratory results from the first hospital day; and a range of clinical data from the first hospital day. Inclusion of a few underreported secondary diagnoses substantially improved the performance of the models, as did the addition of numerical laboratory results. Inclusion of all diagnoses and the full range of clinical data, however, produced little further improvement in the models' statistical performance.

Other data sources also can be merged with administrative data to enrich its content. Addition of aggregate information collected by the decennial census enables researchers to introduce population-level information (e.g., about poverty level, racial distribution) into analyses, as a growing body of evidence links social and environmental determinants with population health (Marmot 2002; Subramanian and Kawachi 2006; Wilkinson and Pickett 2006; Kawachi and Subramanian 2007). Information from the AHA annual survey can be merged with administrative data to describe information about the institutions providing services (e.g., teaching status, bed size, ownership, available services, patient populations). These examples involve merging information about individuals with aggregate, contextual information (e.g., about populations within census tracts, neighborhood characteristics, hospital attributes).

Linkage of two or more data sources containing information on individuals offers considerable additional insight. Most such merges involve Medicare data. Two merged files that link different sets of Medicare data are (1) the CCW and (2) the Medicare Current Beneficiary Survey (MCBS) merged

with Medicare claims. The CCW merges information from different settings of care, as described in the section on Medicare data earlier in this chapter. The merging of MCBS responses with Medicare claims directly links beneficiaries' perceptions (e.g., about their health and health care experiences) with claims indicating their service use. As described on the ResDAC website and at www.cms.gov/mcbs, MCBS is an ongoing longitudinal survey of a representative panel of roughly 12,000 Medicare beneficiaries, including an oversampling of persons less than age 65 and persons aged 85 or older (Iezzoni et al. 2004). Respondents typically remain empanelled in the MCBS for four years and are interviewed in person three times annually. Two versions of the survey (one for respondents who reside in communities and one for respondents who live in institutions) solicit information about physical and sensory functioning, satisfaction with and access to care, out-of-pocket payments, and numerous other topics (Iezzoni 2002). MCBS Cost and Use files merge survey responses with Medicare claims.

Sources external to Medicare also may be merged with Medicare data. For example, NCHS has linked some of its nationally representative surveys to Medicare administrative files. Prior to linking these files, survey respondents must consent; 80 percent typically do, but consent rates vary across surveys (Lillard and Farmer 1997, 694). To protect confidentiality, NCHS carefully regulates the content of these files and access to these data (Iezzoni 2002). As of this writing, selected linked files are available from NCHS (2011a). NCHS surveys linked to Medicare enrollment and claims data for services from 2001 to 2007 include the

- 1994–1998 National Health Interview Survey (NHIS),
- National Health and Nutrition Examination Survey (NHANES) I Epidemiologic Follow-up Study (NHEFS),
- Third NHANES (NHANES III), and
- Second Longitudinal Study of Aging (LSOA II).

NCHS surveys linked to Medicare enrollment and claims data for services from 1999 to 2007 include the

- 1999–2005 NHIS,
- 1999–2004 NHANES, and
- 2004 National Nursing Home Survey (NNHS).

Medicare data are available for analysis through the NCHS Research Data Center. To facilitate the use of these linked Medicare survey data, NCHS has created a public use NCHS-CMS Medicare Feasibility Study data file, which can be downloaded from the NCHS (Centers for Disease Control and Prevention) website. This feasibility file provides selected variables that researchers can

use to determine whether analyses using the linked files would be methodologically sound. Plans are in place for linkage of Medicare Part D data.

Another prominent interagency linkage merges Medicare claims data with the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) program data (Potosky et al. 1993; Warren et al. 2002). SEER gathers information from 17 population-based cancer registries covering about 28 percent of the US population; the number of registries included in SEER continues to grow (www.seer.cancer.gov). Started in 1973, the SEER database contains information on more than 6 million in situ and invasive cancer cases; about 350,000 new cases are added annually. SEER routinely collects data on patients' demographics, primary tumor site, morphology, stage at diagnosis, first course of treatment (followed for up to four months), and vital status. Data elements from SEER (Social Security number, name, sex, and dates of birth and death) facilitate linkage with Medicare claims data. Using a deterministic matching algorithm, the first merge (for patients diagnosed from 1973 to 1989) matched 93.8 percent of persons diagnosed at age 65 or older (Potosky et al. 1993). NCI and CMS update the merge every several years and add Medicare claims during intervening years. To enable comparisons, the database also contains a 5 percent random sample of Medicare beneficiaries not in the SEER registry but residing in SEER areas (i.e., "noncancer cases").

ICD-10-CM and Risk Adjustment

While writing this final section of Chapter 5, I see evidence that people across the US health care delivery system are gearing up for the shift to ICD-10-CM and ICD-10-PCS. Messages from my own hospital's data managers, updates on the NCHS (2011b) and CMS websites, and companies advertising their services to assist health care organizations, large and small, with the transition all anticipate the 2013 start date. The marketing language of some coding consulting firms mimics the rampant hyperbole about potential computer meltdowns at Y2K, urging organizations to get ready for ICD-10 or face the consequences.

Exactly what the new nomenclatures will mean for risk adjustment going forward is unclear. Using crosswalks provided by NCHS that match ICD-9-CM diagnosis codes with their closest ICD-10-CM counterparts, organizations that furnish risk adjustment methods for policy purposes (e.g., DRGs for hospital payment) are adapting their algorithms to the new nomenclature. In 2005, investigators who use administrative data from Australia, Canada, China, Switzerland, the United Kingdom, and the United States met in Banff, Canada, to discuss these implications (De Coster et al. 2006). They identified various research priorities, such as translating the code-based version of the Charlson Comorbidity Index and Elixhauser and colleagues' CCS into ICD-10

and redoing AHRQ's Patient Safety Indicators using ICD-10-based codes (Iezzoni 2010). Some methodological research is already under way, albeit this early work involves ICD-10 rather than ICD-10-CM (Quan et al. 2005; Li et al. 2008). Beyond these activities, researchers will need to develop familiarity with ICD-10-CM and ICD-10-PCS to learn the strengths and limitations of these new classification schemes (Iezzoni 2010).

Background of ICD-10

Starting in the 1980s, WHO followed its mandate to periodically revise the ICD diagnosis nomenclature. WHO and its advisory bodies reviewed proposals for revisions from around the world from 1984 through 1987 and then drafted ICD-10. Delegates from 43 countries gathered in Geneva in 1989 to review the draft ICD-10, and the World Health Assembly approved the final version in May 1990; WHO fully released ICD-10 in 1994. ICD-10 looks different from ICD-9. It has an alphanumeric ordering scheme and new chapter titles, and V and E codes, formerly supplementary listings, are integrated into the main classification. Volume I, the Tabular List, contains the classification at the three- and four-character level and runs to more than 1,000 pages. Underscoring the statistical purpose of the nomenclature and its expanded scope, ICD-10's official title is the *International Statistical Classification of Diseases and Related Health Problems.* In 2002, ICD-10 was published in 42 languages, including the six official languages of WHO. A 2007 version is currently available on the WHO website at http://apps.who.int/classifications/apps/icd/icd10online.

International treaty required the United States to report mortality statistics using ICD-10 in 1999. However, as noted above, ICD-9 and ICD-10 differ. Whereas ICD-9 contained about 5,000 categories, ICD-10 has roughly 8,000. Chapters were added and rearranged. "The shifting of deaths away from some cause-of-death categories and into others resulting from these changes creates discontinuities in cause-of-death trends from 1998, the last year of ICD-9, and 1999, the first year of ICD-10" (Anderson et al. 2001, 5). Whenever it implements a new ICD, NCHS examines the effect of the new classification system on longitudinal trends of death rates by diseases. Using more than 1.8 million death certificates from 1996 for all 50 states and the District of Columbia, analysts recoded cause of death using ICD-10. Preliminary results suggested that the change from ICD-9 to ICD-10 resulted in discontinuities in cause-of-death reporting for certain conditions, including septicemia; influenza and pneumonia; Alzheimer's disease; and nephritis, nephrotic syndrome, and nephrosis.

ICD-10-CM

In September 1994, NCHS and its consultants began reviewing ICD-10 to determine whether a clinical modification was necessary for morbidity reporting in the United States, as it had done with ICD-9-CM (see discussion

earlier in the chapter). Without the clinical modification (CM), 93 countries worldwide are using ICD-10 for morbidity reporting. Some are using it to determine payment levels. Countries using ICD-10 for reimbursement or case-mix adjustment include the United Kingdom (starting in 1995), Nordic countries (Denmark, Finland, Iceland, Norway, and Sweden, starting between 1994 and 1997), France (1997), Australia (1998), Belgium (1999), Germany (2000), and Canada (2001).

Although the reviewers found that ICD-10 was a substantial improvement over ICD-9, they nonetheless advised a clinical modification of the diagnosis nomenclature for American use. NCHS worked closely with a technical advisory panel, physician groups and specialty societies, clinical coders, and others to develop ICD-10-CM. NCHS posted the draft ICD-10-CM on its website for public comment from December 1997 to February 1998, and a new draft ICD-10-CM was completed in May 2002. The new structure allows more expansion to add new codes than did ICD-9-CM. NCHS continually posts updated versions of the nomenclature on its website; a January 2011 update is the most recent (NCHS 2011b).

As was ICD-9-CM, ICD-10-CM is organized into chapters (Exhibit 5.7). However, ICD-10-CM contains approximately 68,000 diagnosis codes, compared with the roughly 13,000 diagnosis codes in ICD-9-CM (Barta et al. 2008). Exhibit 5.8 compares the structure of the two nomenclatures' diagnosis codes. Highlights of changes in ICD-10-CM include (Barta et al. 2008; NCHS 2011b)

- greater specificity of codes;
- expanded detail relating to ambulatory and managed care encounters;
- elimination of V and E code supplemental classifications;
- expanded injury codes, grouped by anatomical site rather than category of injury;
- codes combining information about diagnoses and symptoms, aiming to reduce the number of codes required to fully describe conditions;
- combination codes for poisonings and external causes;
- indicators of laterality (right, left, bilateral);
- inclusion of trimester in obstetrics codes and elimination of fifth digits for episodes of care;
- changes to some time frames specified by certain codes; and
- extensions for episodes of care (e.g., A = initial, D = subsequent, S = sequelae; other extensions are used for fracture codes[13]).

Although ICD-10-CM offers advantages over its predecessor, the dictum "more is better" may not always apply—or may apply only with qualifiers. Certainly, ICD-9-CM had exhausted its potential for expansion, limiting the ability to capture newly recognized diseases and additional clinical details about

EXHIBIT 5.7
ICD-10-CM Table of Contents

the thousands of conditions already coded. However, critical information is still not captured by ICD-10-CM. For instance, ICD-10-CM lung cancer codes do not distinguish adenocarcinoma from small-cell lung cancers, a clinically significant distinction. Assignment of more detailed codes will require additional supporting documentation that busy physicians may resist providing. With increased complexity, the "art of coding" is more likely to come into play, especially in performing such tasks as code sequencing. This potential introduces questions about reliability and the factors that drive coding—complex incentives first recognized in 1981 by Simborg in his warning about "DRG creep."

ICD-10-PCS

Although ICD-9-CM contained procedure codes, ICD-10-CM does not. Instead, in 1995 the Medicare agency (HCFA) hired 3M Health Information Systems (which updates the DRGs) to create a new procedure nomenclature,

Attribute of Code	ICD-9-CM	ICD-10-CM
Number of characters	5	7
First character	Numeric or alpha (E or V)	Alpha*
Second and third characters	Numeric	Numeric
Decimal place location	After first three characters	After first three characters
Additional characters	Fourth and fifth numeric	Fourth, fifth, sixth, and seventh alpha or numeric
Meaning of first three characters	Condition category	Condition category
Characters after decimal place	4 and 5 = etiology, anatomic site, manifestation	4, 5, and 6 = etiology, anatomic site, severity
Seventh digit	N/A	Extension**

EXHIBIT 5.8 Structure of ICD-9-CM and ICD-10-CM Diagnosis Codes

*All letters are used except U.
**Extension represents visit encounter (initial [A], subsequent [D], or sequelae [S]) or sequelae for injuries and external causes.
Source: Adapted from Barta et al. (2008).

ICD-10-PCS (Averill et al. 1998). This classification differs substantially from the procedure categories offered by ICD-9-CM, CPT, and HCPCS. ICD-10-PCS codes are alphanumeric and include seven characters, each of which can have up to 34 possible values (the ten digits 0–9 and the 24 letters A–H, J–N, and P–Z; letters O and I are not used because they could be confused with digits 0 and 1). As shown in Exhibit 5.9, ICD-10-PCS is divided into sections that identify general types of procedures (Averill et al. 2011). While the first character of the procedure code always specifies the section, the second through seventh characters may mean different things in different sections, although they always have the same meaning within sections (Jellish, Zenner, and Goetsch 2007). Different characters indicate such attributes as the general type of procedure, body part, and approach. In the medical and surgical section, for instance, the seven characters have the following meanings:

Character 1 = Section
Character 2 = Body System
Character 3 = Root Operation
Character 4 = Body Part
Character 5 = Approach
Character 6 = Device
Character 7 = Qualifier

EXHIBIT 5.9
ICD-10-PCS Sections

0	Medical and Surgical
1	Obstetrics
2	Placement
3	Administration
4	Measurement and Monitoring
5	Extracorporeal Assistance and Performance
6	Extracorporeal Therapies
7	Osteopathic
8	Other Procedures
9	Chiropractic
B	Imaging
C	Nuclear Medicine
D	Radiation Oncology
F	Physical Rehabilitation and Diagnostic Audiology
G	Mental Health
H	Substance Abuse Treatment

Exhibit 5.10 shows the different body systems assigned as character 2.

The following example clarifies the opportunities and complexities of ICD-10-PCS. To code a coronary angioplasty requires the following approach, starting with the first of the seven digits (Averill et al. 2011, 4):

1. Section: 0 = medical and surgical
2. Body system: 2 = heart and great vessels
3. Root operation: 7 = dilation (expanding an orifice or the lumen of a tubular body part)
4. Body part: 0 = coronary artery, one site; 1 = coronary arteries, two sites; 2 = coronary arteries, three sites; and 4 = coronary arteries, four or more sites
5. Approach: 0 = open; 3 = percutaneous; and 4 = percutaneous endoscopic
6. Device: 4 = drug-eluting intraluminal device; D = intraluminal device; T = radioactive intraluminal device; and Z = no device
7. Qualifier: b = bifurcation; and Z = no qualifier

How easily the coding workforce will adapt to ICD-10-PCS and reliably, efficiently, and accurately assign these procedure codes is unclear. Researchers

also will have to shift their thinking away from the relatively simple four-digit procedural classification system of ICD-9-CM. The detail offered by ICD-10-PCS will likely prove valuable after users pass the initial learning curve.

EXHIBIT 5.10
ICD-10-PCS Medical and Surgical Body Systems

0	Central Nervous System
1	Peripheral Nervous System
2	Heart and Great Vessels
3	Upper Arteries
4	Lower Arteries
5	Upper Veins
6	Lower Veins
7	Lymphatic and Hemic System
8	Eye
9	Ear, Nose, Sinus
B	Respiratory System
C	Mouth and Throat
D	Gastrointestinal System
F	Hepatobiliary System and Pancreas
G	Endocrine System
H	Skin and Breast
J	Subcutaneous Tissue and Fascia
K	Muscles
L	Tendons
M	Bursae and Ligaments
N	Head and Facial Bones
P	Upper Bones
Q	Lower Bones
R	Upper Joints
S	Lower Joints
T	Urinary System
U	Female Reproductive System
V	Male Reproductive System
W	Anatomical Regions, General
X	Anatomical Regions, Upper Extremities
Y	Anatomical Regions, Lower Extremities

Looking Forward

This chapter thus ends on an unsatisfactory note of uncertainty. A potentially transformative change for administrative data lies tantalizingly around the corner, but we do not know how this transition will work out. The implications of ICD-10-CM and ICD-10-PCS for risk adjustment also remain uncertain, although both nomenclatures offer considerable opportunities for enhanced clinical insight that would benefit risk adjustment. Researchers will need to immerse themselves in these classification systems to understand their full potential and employ them to their greatest advantage. The systems' complexity, however, raises additional questions about coding completeness, accuracy, and reliability, especially given pressing financial motivations for code assignments.

The new coding systems will almost certainly challenge all who use them, from coders to researchers studying administrative data. Creative ways to address these complexities must therefore be devised. One possibility, which enterprising designers are probably already developing, is to harness the content of electronic health records to automatically generate diagnosis and procedure codes. Natural language processing (NLP) and other techniques have already been used to extract information about health conditions from electronic records and then populate patients' problem lists (Galanter et al. 2010; Meystre and Haug 2008). An automated problem list created from free-text electronic health records using NLP could improve the efficiency and accuracy of diagnosis coding (Meystre and Haug 2006). Chapter 6 explores these and other possible advantages of electronic health records for risk adjustment.

Notes

1. As of January 2011, HCUP's data set includes hospital discharge information from 40 states: Arizona, Arkansas, California, Colorado, Connecticut, Florida, Georgia, Hawaii, Illinois, Indiana, Iowa, Kansas, Kentucky, Maine, Maryland, Massachusetts, Michigan, Minnesota, Missouri, Nebraska, Nevada, New Hampshire, New Jersey, New York, North Carolina, Ohio, Oklahoma, Oregon, Rhode Island, South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, and Wyoming.
2. Some databases assign the mother's identification number to her newborn infant, making health care claims attributable to the mother and infant difficult to disentangle.
3. The HIPAA Privacy Rule is a national standard that protects individuals' medical records and other personal health information; it applies to

health plans, health care clearinghouses, and certain electronic health care transactions. The rule sets limits and conditions on the use and disclosure of individually identifiable health information without patients' authorization. It also gives individuals certain rights, such as the right to examine and obtain a copy of their health records and request corrections.

4. The Privacy Act of 1974 (5 U.S.C. § 552a) established a code of fair information practices that governs the collection, maintenance, use, and dissemination of individually identifiable information maintained by federal agencies. It prohibits disclosure of information from a federal record system without the written consent of the individual, unless one of 12 statutory exceptions is met.
5. The 21 conditions currently included in the CCW include acute myocardial infarction; Alzheimer's disease; Alzheimer's disease, related disorders, and senile dementia; atrial fibrillation; cancer, colorectal; cancer, endometrial; cancer, female breast; cancer, lung; cancer, prostate; cataract; chronic kidney disease; chronic obstructive pulmonary disease; depression; diabetes; glaucoma; heart failure; hip/pelvic fracture; ischemic heart disease; osteoporosis; rheumatoid arthritis/osteoarthritis; and stroke/transient ischemic attack.
6. Veterans with high household incomes, geographically adjusted, historically have not been eligible unless they met other criteria for a priority group. The income target was relaxed on June 15, 2009. Veterans in the high-income priority group (Group 8) must agree to pay certain copayments to enroll. Information on VA enrollment is available at www.va.gov/opa/publications/benefits_book/benefits_chap01.asp.
7. The following example demonstrates the distinction between principal and primary diagnoses. Suppose an elderly man is admitted routinely for a transurethral prostatectomy (TURP) to treat benign prostatic hypertrophy (BPH). On the day following an uncomplicated TURP, he falls and fractures his hip. He is stabilized and undergoes an open reduction and internal fixation (ORIF) to repair the fracture; postoperatively, he begins rehabilitation. Under UHDDS guidelines, BPH is the principal diagnosis because the patient entered the hospital for BPH treatment. The TURP surgery involved only one or two days, whereas the ORIF required an additional week in hospital. Thus, hip fracture is the primary diagnosis.
8. AHRQ also developed a CCS for Mental Health and Substance Abuse (CCS-MHSA), which was rolled into the CCS for ICD-9-CM diagnoses in 2008.
9. For outpatient surgery, the reason for the surgery is coded as the first-listed diagnosis (reason for the encounter), even if the surgery is cancelled because of a contraindication. For patients admitted for

observation of a medical condition, the medical condition is coded as the first-listed diagnosis.

10. The conditions for which Medicare no longer paid were diagnoses after admission of air embolism, blood incompatibility, catheter-associated urinary tract infection, pressure ulcer, object left in patient during surgery, vascular catheter-associated infection, mediastinitis after coronary artery bypass grafting, and fall from bed (Rosenthal 2007).
11. These reductions will end when CMS starts risk modeling by putting MA utilization rather than FFS utilization into the model.
12. Coding quality was "very good" for heart failure, chronic renal disease, prior coronary bypass surgery, history of pacemaker, complete atrioventricular block, and shock; it was "intermediate" for epilepsy, other cerebrovascular disease, primary or secondary malignancy, and hypertension. This study of AMI patients admitted in 1990 and 1991 also found that 7.6 percent did not meet clinical criteria for AMI and 23.7 percent had a "possible" AMI (most of the possible cases had either chest pain with borderline cardiac enzymes or positive enzymes without chest pain) (Wilson, Smoley, and Werdegar 1996, 14–23).
13. ICD-10-CM fracture codes require a seventh character that indicates whether an initial encounter is for an open fracture or a closed fracture or if a subsequent encounter is for routine healing, delayed healing, nonunion, malunion, or sequelae (Barta et al. 2008). The fracture extensions are: A = initial encounter for closed fracture; B = initial encounter for open fracture; D = subsequent encounter for fracture with routine healing; G = subsequent encounter for fracture with delayed healing; K = subsequent encounter for fracture with nonunion; P = subsequent encounter for fracture with malunion; and S = sequelae.

CHAPTER

6

CLINICAL DATA FROM MEDICAL RECORDS AND HEALTH INFORMATION TECHNOLOGIES

Lisa I. Iezzoni

No other topic in this book has received as much presidential attention as medical records, particularly health information technology (HIT) and electronic health records (EHRs). In his 2004 State of the Union address, President George W. Bush asserted that "by computerizing health records, we can avoid dangerous medical mistakes, reduce costs, and improve care." In 2005, he set a goal to implement universal EHRs within ten years, outlining his plans for increasing HIT use and creating national standards for digitizing, storing, and sharing electronic health information. His successor, President Barack Obama, wasted little time in joining this call. The American Recovery and Reinvestment Act (ARRA) of 2009, the stimulus package enacted shortly after Obama's inauguration, featured HIT prominently among its various provisions. As described later, in its Health Information Technology for Economic and Clinical Health (HITECH) provisions, ARRA established a variety of policy mandates that are setting the stage for EHRs for many years to come (Blumenthal 2009, 1477):

> The HIT components of the stimulus package—collectively labeled HITECH in the law—reflect a shared conviction among the fledgling Obama administration, the Congress, and many health care experts that electronic information systems are essential to improving the health and health care of Americans. However, proponents of HIT expansion face substantial problems. Few U.S. doctors or hospitals—perhaps 17% and 10%, respectively—have even basic EHRs, and there are significant barriers to their adoption and use: their substantial cost, the perceived lack of financial return from investing in them, the technical and logistic challenges involved in installing, maintaining, and updating them, and consumers' and physicians' concerns about the privacy and security of electronic health information. HITECH addresses these obstacles head on, but huge challenges await efforts to implement the law and fulfill President Barack Obama's promise that every American will have the benefit of an EHR by 2014.

This chapter considers the promise and potential challenges of HIT and using EHRs as the source of information for risk adjustment. Stimulated by the ARRA mandates, the HIT field has taken off in the last two years. A

comprehensive review of the current state of electronic information systems in health care is beyond the scope of this text, but it is essential to discuss the major advances that ARRA has made possible and the exciting implications for risk adjustment.

Another burgeoning trend is the involvement of patients in reviewing and annotating their EHRs. Medical records are largely still secondhand, in the sense that clinicians filter all observations; they ask the questions and record patients' answers. Medical records, paper or electronic, thus reflect the realities of human discourse, such as faulty memories, communication failures, subjective judgments, and conscious and unconscious distortions—factors involving both patients and clinicians. Medical records do not always yield a straightforward, complete, and objective account of patients' risks, especially their preferences for care (see Chapter 3). This situation may change in coming years as patients begin documenting their own medical records.

History of Medical Records

Aggregation of notes about individual patients to document their course of illness and care did not begin until the mid-nineteenth century (Siegler 2010). Before then, American hospitals kept records in bound volumes organized chronologically by admissions, not by patients. In conducting rounds, physicians walked from bed to bed, writing brief assessments of patients consecutively by order of encounter. Notations emphasized descriptions of patients and their symptoms and contained scant information on physical findings, diagnoses, or treatment. The following note from Massachusetts General Hospital on August 18, 1824, was typical of the time: "Skin nearly natural, tongue rather dry, with moist red edges. Countenance good. Took hasty pudding at 10 a.m." Other notes contained lengthier descriptions of patients' symptoms and physical findings yet no diagnoses. The following note, also from Massachusetts General Hospital on August 18, 1824, describes a woman of unspecified age:

> Sense of fullness and oppression in the Chest. Dyspnea increased by exercise and by recumbent position. Sense of tightness across thorax. Pains in left side, not constant, but often darting to the epigastric region and to left shoulder. Inability of lying on left side, which brought on great distress and sense of suffocation. Tongue has thin pale coat. . . . Skin now cool and moist, says she has sweat much at times. Thinks she has not lost much flesh, being much swelled.

The colorful description of symptoms is revealing and suggests heart failure with episodes of angina pectoris.

Even in the late nineteenth century, medical record notes remained primarily descriptive. The Mayo brothers in Rochester, Minnesota, generally did not record physical examinations, diagnoses, or treatments. Most of their medical records contained only the date, the patient's age and residence, and descriptions of symptoms, such as "gas on the stomach and poor sleep" and "night terrors—wetting bed" (Clapesattle 1941, 385). The unit medical record that accumulates all information on individual patients in a single file was not fully implemented until 1916 at Presbyterian Hospital in New York City (Kurtz 1943).

Concerns about the quality of medical records arose in the early 1900s. The American College of Surgeons extensively discussed shoddy medical record keeping, and in 1917, its Conference on Hospital Standardization proclaimed, "If good records are kept, it is almost certain that good work will be done" (Hornsby 1917, 7). The Conference asserted that 75 percent of hospitals kept "valueless" medical records because of missing information on histories, physical examinations, and diagnoses. Furthermore, the Conference called the paucity of diagnostic information "premeditated" and "intended to cover up and hide carelessness or incapacity on the part of the surgeon to diagnose the disease" (Hornsby 1917, 7).

Ernest Amory Codman (1917), a founder of the College (Chapter 1), envisioned a heretical new use for medical records: "Our record system should enable us to fix responsibility . . . for the success or failure of each case treated," noting the need for "clear, honest records, no matter how brief, if they fearlessly face the facts." One proposal stipulated that progress notes be written at least once every three days (Ramsey and Kingswood 1923). Raymond Pearl, the statistician to The Johns Hopkins Hospital, advocated standardized printed forms, bemoaning the diverse medical record formats physicians used. As Pearl (1921, 187) observed, "The general scheme or outline which a history is to follow resides, far too often, in the head of the history writer, and there only. And heads, especially of human beings, do vary so!" In 1918, the American College of Surgeons (1918, 2) adopted Codman's vision for medical records:

> Consistent and fearless review of case records by the hospital staff, as here suggested, is a just and effective means to deal with incompetent medical and surgical work in a hospital. Facts are not debatable. . . . If the facts establish evidence that a physician or surgeon is unsafe in judgment, unworthy in character, untrained, lax, lazy, or careless, in all honor and decency that individual should either overcome his deficiencies or withdraw from practice. . . . A wise use of honest case records points the way to great advance in the medical profession.

Despite these exhortations, the medical record did not change significantly for more than four decades. The self-proclaimed independence of

physicians, who railed against any hint of standardization or uniformity, underlay this inertia. Charting changed in the 1960s, when new clinical practices and burgeoning medical knowledge threatened to overwhelm physicians and their record keeping. In the visionary view of Lawrence Weed (1968, 593), it was "necessary to develop a more organized approach to the medical record, a more rational acceptance and use of paramedical personnel and a more positive attitude about the computer in medicine." He proposed organizing documentation around a "problem list," a dynamic inventory of major clinical issues requiring attention. Weed (1971) developed the "SOAP" heuristic for documenting clinical problem lists, organizing information under four major headings:

- *Subjective:* symptoms and subjective feelings reported by patients, capturing patients' views about their health status and health care
- *Objective:* clinical findings, such as physical examinations, laboratory tests, and diagnostic procedure results
- *Assessment:* physicians' conclusions about the status of the problem based on the subjective and objective evidence
- *Plan:* diagnostic, therapeutic, monitoring, and other activities undertaken to address the problem

Weed's problem-oriented medical record and SOAP approach were widely adopted and taught to generations of young physicians.

Despite the explosion of digitized information, the structure and content of written medical records have changed little in the ensuing decades. Weed's early pleas for computerization were largely ignored for many years, except perhaps by laboratories (e.g., for reporting blood test results, albeit on paper printouts). Some patients' medical records still spill into multiple volumes containing hundreds of poorly fastened pages.

Several years after Weed's proposal, Lyons and Payne (1974, 714) reported "questions about how validly medical record information reflects the actual medical practice of the physician rather than only his record-keeping performance." They underscored the central role of medical records in communication among clinicians caring for patients, quoting Avedis Donabedian as saying "the conditions that bring about good care are also responsible for bringing about good recording" (Lyons and Payne 1974, 714). A contrary view held that "classic casebook recording may be a defense for inferior practice and . . . the very best clinician may not use up the ink in one ballpoint pen in his entire career" (Lyons and Payne 1974, 715). Their research results, however, favored Donabedian's presumption: In general, better documentation and better quality are positively correlated. Even the dramatic changes in medical practice over the last 40 years probably have not altered this fundamental relationship.

Going Digital

As noted at the beginning of this chapter, the nation is in the midst of a major transformation of medical records. Although the most obvious change involves the form and format of the records—from paper to EHRs—ARRA envisions something more profound in its "meaningful use" provisions described in the following paragraph. Nonetheless, the United States has a long way to go before institutional and clinical providers nationwide implement EHRs. A 2009 survey of US hospitals found that only 12 percent had adopted either basic or comprehensive EHRs, up from just under 9 percent in 2008 (Jha et al. 2010). Larger, private, and urban hospitals were more likely to have EHRs than were small, public, and rural hospitals. In particular, hospitals caring for a disproportionate number of poor patients are less likely to have EHRs than are other facilities (Jha et al. 2009). Rates of EHR adoption are similarly low among physicians. A national survey of 2,758 physicians conducted in late 2007 and early 2008 found that only 4 percent had extensive, fully functional EHRs and just 13 percent had basic electronic record systems (DesRoches et al. 2008). For both hospitals and physicians, the high costs of implementing HIT pose an enormous barrier to its adoption. Interestingly, a 2007 nationwide survey of home health and hospice agencies found that 41 percent already had EHRs and an additional 15 percent planned to implement electronic records within a year (Bercovitz, Sengupta, and Jamison 2010).

Recognizing the financial impediments to adoption of HIT, the ARRA HITECH provisions committed federal resources to support implementation of EHRs among physicians (Blumenthal 2009, 1478):

> Starting in 2011, physicians can receive extra Medicare payments for the "meaningful use" of a "certified" EHR that can exchange data with other parts of the health care system. These payments can total as much as $18,000 in the first year in the case of physicians who adopt in 2011 or 2012, with at least $15,000 for physicians who adopt in 2013 and a slightly lower amount for those who do so in 2014; incentives are gradually reduced and then ended in 2016. Thus, physicians demonstrating meaningful use starting in 2011 could collect $44,000 over 5 years. Waiting until 2013 would result in a maximum bonus of $27,000 over 3 years. Experts estimate the cost of purchasing, installing, and implementing an electronic-records system in a medical office at about $40,000.

Whether these resources will be sufficient to spur widespread adoption of electronic record systems remains unclear. A national survey found that 51.8 percent of office-based physicians planned to apply for the EHR financial incentives in 2011, up from 41.1 percent in 2010 (Hsiao et al. 2011, 3).

Questions persist about whether physicians and hospitals largely serving impoverished populations will have sufficient resources to implement HIT; the funds that Congress has directed state Medicaid agencies to channel to them might be insufficient or might be diverted to other purposes.

Implementing EHRs in small physician practices may be particularly challenging (Felt-Lisk et al. 2010). An especially difficult problem for small practices is technological support when they experience difficulties with the EHRs (they do not have the in-house support a hospital would have) or when HIT system capabilities do not fit physicians' needs (e.g., EHRs may be organized differently than physicians' paper records or may not feature the search capabilities they want).

The meaningful use objectives (Exhibit 6.1) are to make HIT more than a means of digitizing health information. The goal is to use the information to improve health care quality, reduce medical errors, and help clinicians make better decisions (Blumenthal 2010; Blumenthal and Tavenner 2010). "Meaningful use, coupled with large financial incentives, may signal the beginning of the end of health care as a cottage industry" (Jha 2010, 1710). Congress structured the meaningful use provisions to evolve over time and to eventually reward providers for using EHRs to improve processes of care and outcomes. In 2009, only 2 percent of US hospitals reported having EHRs that would meet the federal meaningful use criteria (Jha et al. 2010). Newly developed Regional Health Information Technology Extension Centers, which provide technical assistance, guidance, and information on best practices for HIT, intend to assist providers in implementing electronic information systems (Torda, Han, and Scholle 2010).

EXHIBIT 6.1
Federal Meaningful Use Objectives for Electronic Health Records

All eligible professionals and hospitals must achieve the following set of core EHR objectives to qualify for incentive payments:

- Record patient demographics (sex, race, ethnicity, date of birth, preferred language, and in the case of hospitals, date and preliminary cause in the event of death)
- Record vital signs and chart changes (height, weight, blood pressure, body-mass index, growth charts for children)
- Maintain up-to-date problem list of current and active diagnoses
- Maintain active medication list
- Maintain active medication allergy list
- Record smoking status for patients aged 13 or older
- Individual professionals: provide patients with clinical summaries for each office visit; hospitals: provide an electronic copy of hospital discharge instructions on request
- On request, provide patients with an electronic copy of their health information (including diagnostic-test results, problem list, medication lists, medication allergies, hospital discharge summary, and procedure information)

EXHIBIT 6.1 (Continued)

- Generate and transmit permissible prescriptions electronically (does not apply to hospitals)
- Implement computerized provider order entry (CPOE) for medication orders
- Implement drug–drug and drug–allergy interaction checks
- Implement capability to electronically exchange key clinical information among providers and patient-authorized entities
- Implement one clinical decision support rule and ability to track compliance with the rule
- Implement systems to protect privacy and security of patient data in the EHR
- Report clinical quality measures to the Centers for Medicare & Medicaid Services or the state

Eligible professionals and hospitals may select any five choices from the following list:

- Implement drug formulary checks
- Incorporate clinical laboratory test results into EHRs as structured data
- Generate lists of patients by specific conditions to use for quality improvement, reduction of disparities, research, or outreach
- Use EHR technology to identify patient-specific education resources and provide those resources to the patient as appropriate
- Perform medication reconciliation between care settings
- Provide summary of care record for patients referred or transitioned to another provider or setting
- Submit electronic immunization data to immunization registries or immunization information systems
- Submit electronic syndrome surveillance data to public health agencies

Additional choices for hospitals:

- Record advance directives for patients aged 65 or older
- Submit electronic data on reportable laboratory results to public health agencies

Additional choices for eligible professionals:

- Send reminders to patients (per patient preference) for preventive and follow-up care
- Provide patients with timely electronic access to their health information (including laboratory results, problem list, medication lists, and medication allergies)

Source: Adapted from Blumenthal and Tavenner (2010, 502–503).

Whether EHRs improve care remains an open question. DesRoches and colleagues (2010) looked at associations between EHR adoption and various quality metrics and found limited evidence that EHRs improve quality. Another study found that HIT benefits to health care quality primarily accrue to academic medical centers rather than to other types of hospitals (McCullough et al. 2010). Computerized provider order entry (CPOE) is included among the required meaningful use criteria in expectation that, when combined with decision support tools, it will reduce medication errors and improve the efficiency of care processes in hospitals. One study of 62 hospitals, however, found that this promise may not be achieved with current technologies. Using simulation techniques, Metzger and colleagues (2010) found that the decision support tools identified only 53 percent of medication orders that would have caused fatalities and 10 percent to 82 percent of orders that would have precipitated serious adverse drug events. Similarly, a statewide survey of physicians in Massachusetts in 2005 looked at associations between EHR use and various quality indicators (Poon et al. 2010). The investigators failed to find a direct association between EHR use and superior quality; however, they did find that physicians using EHRs scored better on certain performance measures than did physicians not using EHRs. In particular, the strongest associations involved the problem list, visit note, and radiology test result components of the EHR and performance measures relating to women's health, colon cancer screening, and cancer prevention.

The expectation is that quality measurement studies will be much easier to conduct using EHRs versus paper records. However, this supposition may not always be the case. One study reviewed the data required to identify the indicators used in the RAND Quality Assessment Tools system (Roth et al. 2009) and found that only roughly one-third of the indicators were readily accessible from EHRs. Different EHR products have different formats for recording the same data elements, making it hard to collect information in a standardized way. In particular, clinical variables, such as social history or disease-specific histories, change frequently and thus must be periodically updated in medical records. Some EHRs, however, automatically retain the information in these fields from visit to visit. Consequently, having these data fields already filled may make clinicians less likely to update the information. Another problem involves the practice of copying and pasting information from EHRs from one encounter to the records of a subsequent encounter. Especially in teaching hospitals, this practice may perpetuate errors in documentation.

Given the current interest in patient-centered medical homes, another question is whether the meaningful use criteria are sufficient to support the needs of a fully functioning medical home. EHRs are considered central to the concept of medical homes, but "even today's leading electronic health records do not include much of the functionality that will be required to transform the care of chronically ill patients" (Bates and Bitton 2010, 614).

Several domains of HIT require further development to fully serve patient-centered medical homes: clinical decision support (information systems that improve decision making), team care support (systems that improve communication among clinical specialists caring for patients), support for care transitions (i.e., movement from one care setting to another), personal health records for use by patients or their proxies, telehealth technologies, and measurement capabilities for improving care. Without these capabilities, EHRs may prove no better than paper records at supporting medical homes and may cause difficulties. Two physicians described problems with implementing EHRs in a large, multidisciplinary medical home in southern New Jersey; limitations in the EHR technology they deployed led to medication errors, workflow interruptions (when the software became sluggish and unreliable), and a variety of other problems that often dog paper medical records. In particular (Fernandopulle and Patel 2010, 624):

> As in many other practices that have implemented electronic records, we found that documentation time required for physicians actually increased compared to paper charting. We found that the system of point-and-click templates, designed to facilitate quick notes, was inadequate for the variability of patients in our practice. The notes generated by such methods yielded documents that were difficult to read for future reference.

While investments in HIT have remained low in the private sector, one exemplar of EHR adoption is the Department of Veterans Affairs (VA), which for more than two decades has worked assiduously to develop computerized patient records, electronic radiology imaging systems, and electronic systems for laboratories and medication orders (Byrne et al. 2010). Known together as the Veterans Health Information Systems and Technology Architecture (VistA), these electronic systems aim to improve patient outcomes and increase efficiency in VA hospitals. One estimate suggests that the VA has achieved cumulative benefits of $3.09 billion over its total costs of investment in the VistA system (Byrne et al. 2010). These savings have accrued from a variety of improvements. One example is the bar-code medication administration system implemented in 1998. This system validates the administration of unit dose and IV medications in real time at the point of care. Cost savings come from reduced inpatient medication errors prevented by this bar-code system. Another example is the electronic system for picture archiving and communication in radiology, which has saved money by reducing film supply costs, film processor maintenance costs, time spent on film processing by radiology department clerks, and floor space costs for the film library (Byrne et al. 2010).

Clinical information from the medical care process tremendously improves the medical meaningfulness of risk adjustment. The question today is whether going digital will further improve the quality and utility of that information. In addition, as described later in this chapter, the cost of extracting

information from paper records is tremendous, whereas large data sets can be produced by extracting clinical information from electronic systems and used to develop clinically based risk adjustment methods at minimal expense. Keep in mind that health information systems are in flux, and what I write today may be long out of date when you read this chapter.

Reliability of Clinical Information

Risk adjustment is only as valid and reliable as its underlying data. As described in Chapter 5, many structural factors raise concerns about the quality of administrative data, such as limitations of the coding scheme and financial incentives. The quality of clinical data also generates questions, regardless of whether their format is digital or paper. As described later in that chapter, for example, within several years of publication of New York State reports about death rates after coronary artery bypass grafting (CABG) surgery, critics charged that some cardiothoracic surgeons exaggerated their patients' risks, such as the extent of lung and heart disease. In clinical settings, the subjectivity of many medical findings hampers efforts to detect gaming and manipulation.

Blurry boundaries between medical art and science (i.e., subjective and objective judgments) require constant vigilance. Interobserver variation in critical clinical findings can significantly and negatively affect patient care (Elmore and Feinstein 1992, 567):

> For the many phenomena in clinical medicine that are not easily measured and that lack a convenient or definitive "gold standard," the consequences of observer variability may be dramatic. A radiologist's interpretation of a mass lesion may lead to an expensive and invasive diagnostic work-up; a pathologist's reading of a tissue slide may determine whether a woman keeps her breast or loses it; a research technician's decision about primary clinical data may affect the final results of a research project.

The medical literature amply documents physicians disagreeing among themselves about the presence of a physical finding and the interpretation of common tests (e.g., chest radiographs, mammograms, electrocardiograms [ECGs]) and sophisticated studies (e.g., endoscopy, angiography, radionuclide imaging). Research has also found tremendous intraphysician variability: physicians disagreeing with themselves on rereview of a physical examination or a test. Physicians have mobilized to understand better the implications of and solutions for unreliable or questionable physical examination findings. The Society for General Internal Medicine sponsors a Clinical Examination Research and Education Interest Group devoted to this concern. The American Medical Association (AMA) has taken on this issue through its major publication, the *Journal of the American Medical Association* (*JAMA*). *JAMA* publishes a feature called "The Rational Clinical Exam,"

in which authors evaluate the quality of the evidence documenting various clinical conditions. In one such article, "Is This Patient Having a Myocardial Infarction?," Panju and colleagues (1998) found high precision of physical examinations for dyspnea and displaced cardiac apex beat but low precision for pulmonary rales and hepatomegaly. Precision of ECG interpretation depended on the physicians' experience level: Residents disagreed on 70 percent of determinations about whether ECGs showed a heart attack, while cardiologists disagreed only 10 to 23 percent of the time. The following titles are examples of recent articles in this series:

- "Does This Patient Have a Pleural Effusion?" (Wong, Holroyd-Leduc, and Straus 2009)
- "Does This Patient with Palpitations Have a Cardiac Arrhythmia?" (Thavendiranathan et al. 2009)
- "Does This Patient Have Delirium?" (Wong et al. 2010)

A *JAMA* website (http://jamaevidence.com/resource/523) and other publications provide information for use in evaluating the accuracy and reliability of clinical examinations (Sackett 1992; Simel and Rennie 2009).

Questions about interrater reliability and the relative value of different clinical findings should help guide choices of candidate risk factors. Especially in observational studies, researchers generally cannot confirm independently whether clinical risk factor information is accurate or reliable. Even if researchers could rereview or reinterpret critical tests or findings (e.g., radiographs, ECGs), such evaluation might still be challenging. Without explicit review criteria, researchers are equally susceptible to bias or interobserver variation. When possible, analysts conducting rereviews should be blinded to patients' outcomes or other critical factors that might bias their interpretations. This strategy pertains only to clinical findings based on "hard" data that can be retrieved and reviewed retrospectively (e.g., radiographic images, pathology tissue slides, ECG tracings). Rereviews are typically not possible for risk factors relating to findings that are ephemeral, are transitory, or require patients to be available for reexamination (e.g., physical findings).

The circa 1990 version of the Computerized Severity Index (later renamed the Comprehensive Severity Index or CSI) offers lessons on using clinical data in risk adjustment. The CSI relied heavily on detailed clinical information. Roughly half of its clinical factors were numeric values, such as vital signs or laboratory findings, whereas others were largely descriptive and qualitative (Iezzoni and Daley 1992). Many factors represented findings that are often measured unreliably, such as jugular venous distention, Kussmaul respiration, rales, and pulsus paradoxus. Other items required patients' reports and thus depended on recall or patients' thresholds for pain and complaints about pain, acuity of self-observation, and willingness to reveal personal facts. Examples included nausea, anorexia, numbness, tingling, and

severe headache. Most differences in patients' reports of these factors probably vary randomly across providers. Nevertheless, systematic differences relating to language, culture, education, and socioeconomic status could occur across patient populations in the detail and nature of such self-reports.

The CSI example illustrates the considerations raised by using potentially unreliable clinical findings for risk adjustment. Regardless of whether certain clinical findings are unreliable, physicians typically use these observations to make therapeutic and prognostic judgments about patients, as indicated by the earlier discussion of *JAMA*'s "Rational Clinical Exam" feature. Therefore, disregarding potentially subjective findings in assessing patients' risk is problematic, especially if the risk-adjusted information must be viewed by clinicians as medically meaningful and thus sufficiently valid to motivate their behaviors (e.g., in quality improvement initiatives). The CSI's developers therefore chose to include such potentially unreliable but common clinical findings among their sets of diagnosis-specific risk factors. Nonetheless, when using such potentially unreliable findings as risk factors, researchers must remember their inherent subjectivity. This subjectivity raises the critical question of whether a potential for bias exists: Could systematic factors bias assessments of these clinical findings by affecting the subjective interpretations of clinicians or reports of patients?

The answer to this question depends on the purpose. "The New York State Cardiac Surgery Reporting System (CSRS) is arguably the gold standard for the public reporting of hospital and physician quality" (Jha and Epstein 2006, 844). Nonetheless, as discussed later in this chapter, public reporting of risk-adjusted CABG death rates might motivate cardiothoracic surgeons in New York State, at least theoretically, to exaggerate reports of risk factors requiring subjective interpretation to make their patients appear more severely ill than warranted by the patients' actual conditions (Green and Wintfeld 1995). A potentially subjective risk factor is the New York Heart Association functional classification for cardiovascular disease patients, which reflects how cardiac symptoms impede usual activities and is based on both observation and patients' reports. Thus, two opportunities exist—from both clinicians and patients—for bias to compromise accuracy.

Quality of Medical Records and Other Risk Factor Documentation

> The clarity and logic of medical recording is a direct reflection of the clarity and logic (or lack thereof) of the medical practitioner's thoughts. An ambiguous, illegible chart with a confusion of orders and counter-orders, incomplete physical examinations, and absence of clear impressions or plans most likely reflects the same confusion in that patient's physician. (Bradbury 1990, 25)

As noted earlier, the completeness, accuracy, and validity of medical records have raised concerns from the outset of record keeping. Physicians' indecipherable handwriting, a long-standing joke, has serious consequences. Illegible handwriting is associated with medical errors, especially erroneous medication prescriptions (Boehringer et al. 2007). Photocopied records used for abstraction outside hospitals or physicians' offices are even less legible than the originals. In a classic study comparing different risk adjustment methods, Thomas and Ashcraft (1989, 487) could not abstract 40 percent of their 431 charts using MedisGroups, which required detailed information from physicians' notes. Photocopied charts made key elements in the clinical record unreadable. EHRs largely eliminate problems caused by illegibility of medical documentation.

Relatively little has been published on the completeness and accuracy of medical records. This limited and perhaps dated literature suggests that medical records suffice for certain purposes but that records frequently fail to completely or accurately represent certain risk factors and the care patients receive (Romm and Putnam 1981; Zuckerman et al. 1975). Feigl and colleagues (1988) compared inpatient and radiation therapy records of patients diagnosed with certain cancers. While inpatient records amply documented laboratory and other diagnostic test results, consultations, and inpatient treatments, important information from outpatient settings about the extent of the cancer (e.g., mammography results) was poorly documented. A comparison of medical records with patients' self-reports found that 83 percent of cases demonstrated at least one error in medication documentation (Beers, Munekata, and Storrie 1990). A detailed comparison of paper records and EHRs at eight primary care practices in Ohio found that EHRs documented certain administrative tasks better than did paper records (e.g., 100 percent of EHRs versus 86 percent of paper records contained required clinician signatures) (Elder et al. 2010). However, neither record type thoroughly documented various important aspects of care: 73 percent of EHRs documented clinical interpretations of test findings versus 64 percent of paper records, and 80 percent of EHRs documented patient notification of test results versus 66 percent of paper records.

Some information may be sensitive and pose risks to patients if it is disclosed, and therefore such data are excluded from medical records. A major example is information about genetic testing. Despite the 2008 Genetic Information Nondiscrimination Act, which protects genetic test results, the law does not prevent discrimination based on the use of this information for life, disability, or long-term care insurance purposes (Klitzman 2010). The AMA Code of Medical Ethics indicates that physicians may need to maintain separate files for genetic testing results to make sure that the results are not sent to health or other insurance companies inappropriately.

One study observed hundreds of patient visits to a primary care pediatric clinic and doctors' (interns and residents) use of free-text medical records versus structured encounter forms (Duggan, Starfield, and DeAngelis 1990). When physicians used the structured form, they recorded significantly more information (overall care, history, physical examination, developmental assessment, guidance, and follow-up). However, these physicians also documented a more comprehensive physical examination than was observed. Therefore, at least for examinations, "use of the structured form exerted a greater impact on recording than on performance" (Duggan, Starfield, and DeAngelis 1990, 111).

Potential Biases in Medical Record Information

Various factors can bias medical record documentation. My colleagues and I conducted a study in which trained nurse researchers reviewed more than 1,000 records of elderly Medicare beneficiaries using the Complications Screening Program (CSP) (Iezzoni et al. 1999). Some of the patterns in the CSP findings surprised us and raised questions about potential biases in medical record documentation. When quality of care or other complications occurred during a hospitalization, the physicians, nurses, and other clinicians documented all activities and findings more thoroughly than they did in the absence of complications. Even routine monitoring was documented more thoroughly in the charts of patients with complications.

Completeness of documentation may vary by type of provider. In a study comparing teaching and nonteaching hospitals, my colleagues and I found that physicians' notes at nonteaching hospitals generally focused exclusively on the acute presenting illness (Iezzoni et al. 1989, 1990). We hypothesized that nonteaching hospital physicians did not document chronic coexisting diseases because these physicians generally were the patients' primary providers and had detailed knowledge of their histories from prior encounters. Furthermore, these physicians generally supervised all inpatient decisions and thus did not need to inform other physicians about the patients' histories through extensive notations. In contrast, at teaching hospitals, physicians' notes contained copious information on both acute and chronic conditions. These documentation differences likely reflected role differences between teaching and nonteaching physicians. At teaching hospitals, interns and residents unfamiliar with the patient generally wrote the admission note, which therefore contained information extracted during a comprehensive review of the patient's clinical situation. Furthermore, as interns and residents extensively cross-cover each other during weekends and evenings, the medical record is the vehicle for transmitting critical information to covering physicians.

Records at both teaching and nonteaching hospitals contained scant information on functional status and social history. Most functional and

social information, when it was recorded, appeared in nurses' notes (Iezzoni et al. 1990). We had initially intended to create a measure of patients' living situations (e.g., whether they lived alone or with others) but could not do so because of missing documentation. Other investigators confirm the absence of social history information in medical records (McDiarmid, Bonanni, and Finocchiaro 1991; Mansfield 1986), although this situation may change if hospitals implement EHR templates that include slots for documenting social histories. Another study of inpatient records found inadequate documentation of elderly patients' functional deficits not only in physicians' notes but also in medical record entries by nurses and physical and occupational therapists (Bogardus et al. 2001).

At teaching hospitals, trainees at multiple levels (from interns to subspecialty fellows) often write copious notes. Various attending physicians (e.g., patients' primary care physicians, specialist consultants), nurses, therapists, pharmacists, social workers, and others also document medical records. This mix of caregivers frequently produces inconsistent information about specific risk factors. Furthermore, data collection guidelines may affect risk factor information. For instance, the 1990s version of MedisGroups required that reviewers abstract most information from physicians' notes. Vital signs were among the data not abstracted from physicians' notes. While this strategy speeds chart reviews, it could affect comparisons across hospitals that have different levels of physician coverage. The records of patients at tertiary teaching hospitals with round-the-clock teams of physicians may contain multiple notes, whereas the records of similar patients at small community hospitals may contain only one daily physician note. Documentation differences between hospitals with versus without constant physician coverage are particularly relevant to clinical findings that wax and wane, such as wheezing, pericardial friction rubs, or S3 gallop heart sounds.

The downside of having multiple physicians document patients' records is the need to deal with the inevitable discrepancies in their clinical findings as documented in their notes. My colleagues and I encountered this difficulty in the early 1990s when we hired research nurses to review medical records using MedisGroups (Iezzoni et al. 1992a). The medical records came from complicated patients at tertiary teaching hospitals. First, the sheer volume of many of the records undermined the reliability of data abstraction; patients with complicated conditions underwent numerous procedures and had numerous clinician encounters, and the nurse reviewers frequently missed a particular notation or laboratory report. Second, the MedisGroups vendor instructed the reviewers to take clinical information from the notes of the most senior physician (i.e., attending physicians' notes took precedence over those of interns). Not unexpectedly, they encountered instances in which senior specialists from different disciplines disagreed about their clinical observations. These occurrences compromised our ability to meet the designated

MedisGroups data collection target of 95 percent accuracy. EHRs will not solve these types of problems.

Biases in the extent of documentation of risk factors across hospitals could result directly from critical differences in practice patterns, including variations in use of diagnostic tests or comfort-measures-only orders. Non-random variations in practice patterns could compromise the ability to produce unbiased assessments of patients' risks across providers (e.g., when physiological values are not measured, they cannot contribute to risk assessment). The story recounted in Chapter 3 about the Pennsylvania hospital with high death rates among apparently mildly ill cancer patients exemplifies this problem. At the time, all Pennsylvania hospitals were required to report MedisGroups information to risk-adjust their mortality rates; MedisGroups (beginning in the mid-1980s) relied on extensive key clinical findings (KCFs) abstracted from medical records. When oncologists at this hospital investigated the worrisome deaths, they found that they involved patients, often with widely metastatic disease, who desired comfort measures only, eschewing even routine blood draws for serum chemistry (e.g., sodium, potassium) and hematology testing (e.g., hematocrit, white blood cell counts). No physiological KCFs were available to capture patients' true risks of imminent death.

In some situations (e.g., when risk-adjusting CABG mortality rates), reliance on technologically sophisticated tests for specifying risk factors is reasonable: With few exceptions, all CABG patients undergo extensive evaluations before surgery. However, the results of sophisticated tests could potentially bias estimates of risks across providers with different technological capabilities. For instance, if certain tests are performed only at teaching hospitals, information from those tests may bias comparisons across teaching and nonteaching facilities. Furthermore, because most preoperative testing for elective surgeries now occurs in outpatient settings, a major challenge is linking information from outpatient sources to information from the inpatient stay.

Effects of Published Performance Reports on Data Quality

External forces, most notably publication of risk-adjusted outcome rates, could potentially bias reports of clinical risk factors. In regions that publish performance report cards, some observers have speculated that clinicians perform more tests to identify more risk factors (i.e., to make their patients look sicker). Although this possibility is unproven, it nevertheless heightens concerns about inappropriate services and, paradoxically, increasing costs, especially in competitive health care marketplaces. A more subtle response to public reporting is coloring; in other words, clinicians may color—distort or manipulate—the risk factors they document.

Manipulation of clinical information is certainly possible. With today's powerful pharmacological and technological interventions, physiological parameters can be "normalized" or modulated in gravely ill patients. Differences in

practice patterns relating to use of these sophisticated interventions could affect comparisons of ICU performance: "'Aggressive' treatment may mask extreme changes in physiology that would be identified in another ICU that waits and reacts to changes in patients' physiology" (Teres and Lemeshow 1994, 95). Some manipulation in reporting physiological values could be subliminal or reflect legitimate uncertainty characteristic throughout clinical practice. For instance, if during a manual blood pressure reading the Korotkoff sounds disappear between 104 and 106 mmHg, knowing that a diastolic pressure of 105 mmHg or greater results in a higher risk rating could affect which value is recorded (i.e., 104 or 106).

Whether such documentation shifts occur—and whether they reflect fact or fiction—is difficult to assess (Shahian et al. 2001). The best-studied example involves the report cards on CABG mortality rates that New York State has published for more than two decades. Clinicians consulting with New York's Department of Health had specified the CABG risk factors, including low cardiac ejection fraction, stenosis of the left main coronary artery, unstable angina, New York Heart Association functional class, and various chronic illnesses (Hannan et al. 1990). Physician leaders therefore felt assured of the clinical credibility of the risk adjustment models. In 1991, however, New York *Newsday* published information on CABG volumes and mortality by individual cardiothoracic surgeon after winning a Freedom of Information Act lawsuit, and a public furor ensued. Cardiothoracic surgeons warned that they might refuse to operate on patients who were severely ill and had a lower likelihood of surviving surgery, worried that the deaths of these patients would leave a "black mark" against their name. As a result, family members of extremely sick patients reported they were unable to find a surgeon to operate on their loved ones.

Green and Wintfeld (1995, 1230–31) analyzed the prevalence of risk factors reported quarterly around the public releases of New York's CABG mortality reports and found that "the reported prevalence of chronic obstructive pulmonary disease, unstable angina, and low ejection fraction increased sharply in the first quarter of 1990, just after the first mortality report had been distributed to hospitals. In the years 1989, 1990, and 1991, reported prevalence of chronic obstructive pulmonary disease was 6.9 percent, 12.4 percent, and 17.4 percent, respectively; corresponding prevalence of unstable angina was 14.9 percent, 21.1 percent, and 21.8 percent." Some reporting increased as a result of changes in how risk factors were defined. Nonetheless, some findings suggested that documentation patterns might have changed, at least at certain hospitals, since public reporting began (Green and Wintfeld 1995, 1231):

> Patterns in the data suggested that some institutions outpaced the overall trend toward an increased reporting of risk factors. For example, at one hospital, the

> reported prevalence of chronic obstructive pulmonary disease increased from 1.8 to 52.9 percent; at another hospital, the reported prevalence of unstable angina increased from 1.9 to 20.8 percent. Variation in the reporting of some risk factors was far greater each year than could reasonably be attributed to differences in case mix. For example, in 1991 the prevalence of chronic obstructive pulmonary disease reported by surgeons ranged from 1.4 to 60.6 percent, and for unstable angina, the range was 0.7 to 61.4 percent.

New York State officials were also concerned about shifts in reporting of risk factors. To monitor and guard against this practice, they instituted independent audits and requests that hospitals revise problematic documentation. Nevertheless, "there will always be some room for judgment . . . in determining whether a patient has certain risk factors" (Chassin, Hannan, and DeBuono 1996, 396). As noted earlier, some risk factors are inherently subjective (e.g., New York Heart Association functional class). Comprehensive evaluation of the accuracy of such factors would require examinations of patients by independent, unbiased doctors because simply rereviewing medical records would not suffice. Mark R. Chassin, MD (2002, 47–48), who oversaw the initial years of New York's CABG report cards, remained convinced that the CABG risk adjustment method was sufficiently accurate for public reporting:

> Of course, the acid test for whether the risk factor data are accurate or not is how well the logistic regression model that is used for risk adjustment works to predict mortality. On this score, the model has proved extremely robust over the years, with C-statistics averaging just over 0.8 and tests of calibration demonstrating accuracy of predictions over all levels of predicted mortality.

Risk Adjustment in Clinically Enriched Electronic Databases

Extensive clinical information from electronic data systems should provide rich information for risk adjustment and improve its predictive ability beyond that of administrative data—at least that is the hope and expectation. Technical problems, such as "interoperability" difficulties (i.e., incompatibilities among different HIT systems) challenge linkage of digital health information across diverse sources. Restrictions on data sharing imposed by the Health Insurance Portability and Accountability Act (HIPAA) are another problem. A detailed description of HIPAA privacy provisions is beyond the scope of this text (Terry 2010; Wilder 2010); suffice to say they raise significant logistical and feasibility concerns about creating large health databases. Kelly (2010) described the difficulties involved in building a prototype for the Nationwide Health Information Network, in which he and his colleagues tried to aggregate data from 15 organizations in four states. They not only confronted technological difficulties relating to interoperability problems but

also had to get permission, per HIPAA provisions, from every patient whose information would be included in this massive data set. Despite these challenges, data sharing among organizations around the country is increasing rapidly. In the eyes of one key participant, the chief information officer of the massive Kaiser Permanente health care system, data sharing needs to meet several criteria to be feasible and widely used: It must be fast, cheap, sustainable, high quality, and safe (La Chance 2010).

Early efforts suggest that clinically enriched electronic databases could substantially improve risk adjustment models. A study from 22 St. Louis hospitals found that predictions of in-hospital mortality improved substantially when clinical laboratory values were added to diagnosis codes from administrative data (Pine, Jones, and Lou 1998). A study of almost 5,000 dialysis patients found that models produced significantly better predictions using detailed clinical data than they did using administrative information (Mesler et al. 1999). Similarly, Hannan and colleagues (1992), the academic team that oversaw New York's CABG modeling effort, compared the ability of two risk adjustment models to predict in-hospital CABG deaths, one using an administrative file derived from the New York Statewide Planning and Research Cooperative Systems and the other using information from a clinical database compiled through the Cardiac Surgery Reporting System. The latter contained detailed CABG risk factors, whereas the former included only standard administrative information. The model that used information from the clinical database was significantly better able to discriminate between persons who lived and died, although the addition of just three of the Cardiac Surgery Reporting System's risk factors (reoperation, ejection fraction, and greater than 90 percent narrowing of the left main trunk) significantly improved the predictive ability of the other model.

Researchers in California also looked at the ability of models using clinical versus administrative data to predict hospital mortality rates (Parker et al. 2006). They linked 38,230 clinical data records from the California CABG Mortality Reporting Program of patients who underwent surgery in 2000 and 2001 to records in the California patient discharge data abstract. They used the present on admission (POA) indicator for each discharge diagnosis in the abstract to develop indicators of risk factors present on admission rather than occurring later during the hospital stay (e.g., as complications of care). Compared with the clinical data, the administrative data showed lower prevalence and risk factors. The clinical model for predicting mortality had a somewhat higher c-statistic (0.82) than did the administrative model (0.80). The POA indicator may have enabled California's administrative model to outperform administrative models that do not use that information. The clinical model found that 17 hospitals were outliers, while the administrative model tagged 16 hospitals as outliers; 12 of these hospitals were deemed outliers by both models. Therefore, slightly more than one-

third of California's hospitals would be viewed differently by models using clinical versus administrative data. A study of CABG outcomes in Massachusetts, which also compared results using administrative versus clinical data, found considerable discrepancies between the two (Shahian et al. 2007). At the time, Massachusetts did not yet have POA indicators for diagnoses.

Research suggests, however, that adding a substantial number of clinical risk factors to administrative data (as opposed to just a few) may not significantly improve predictions, especially if administrative data are supplemented by POA codes (Pine et al. 2007, 2009a; Fry et al. 2007). One study explored this question using information from discharge abstracts from 2000 to 2003 along with detailed clinical information (MedisGroups data) to predict inpatient mortality at 188 Pennsylvania hospitals (Pine et al. 2007).[1] Across the five conditions and three surgeries studied, the mean c-statistic for the model using administrative data was 0.79. Sequentially adding POA codes and then numeric laboratory data collected at the time of admission substantially improved the model's predictive ability (to c-statistics = 0.84 and 0.86, respectively). The additional inclusion of more complex clinical data, such as vital signs, blood culture results, and other key clinical findings, improved the model's predictive ability only marginally (to 0.88). Given that these clinical data were extracted by hand, they were expensive to obtain, raising questions about the value of adding the clinical information versus costs. As clinical data become available for download, the costs of adding these data should become minimal.

Tabak, Johannes, and Silber (2007) developed predictive models for in-hospital mortality using discharge abstract data, automated laboratory data, and manually collected data (vital signs and altered mental status). They used information from 194,903 discharges from 2000 to 2003 across 71 Pennsylvania hospitals that routinely gathered clinical information for risk adjustment.[2] The laboratory findings contributed more substantially to predicting deaths than did any other risk factors, with the exception of level of consciousness for predicting stroke deaths. Tabak and colleagues (2010) extended this work, using data from 1,271,663 discharges from 2000 to 2001, including demographics, laboratory findings on admission, ICD-9-CM principal diagnosis codes, and secondary diagnoses codes, to develop 39 disease-specific automated clinical models.[3] They then added to these automated models information manually abstracted from medical records: systolic blood pressure, diastolic blood pressure, respiration, heart rate, temperature, and altered mental status (Glasgow Coma Scale score or notation indicating disorientation, stupor, or coma). Common important predictor variables included age, albumin, blood urea nitrogen or creatinine, arterial pH, white blood cell count, glucose, sodium, hemoglobin, and metastatic cancer. The average c-statistic for the clinical models using automated data was 0.83. Addition of the manually abstracted clinical information increased the average c-statistic only to 0.85.

The researchers concluded that, overall, the automated laboratory results made the highest relative contribution to predicting inpatient mortality.

A clear goal of using EHRs for risk adjustment is to eliminate the need for manual data abstraction. Obtaining numeric laboratory or other data from automated systems that store the data in numeric format is relatively straightforward. More complex is gathering data from narrative texts, such as notes written by clinicians. Considerable effort has gone into producing coding schemes to capture this type of information. The Unified Medical Language System (UMLS), created and maintained by the National Library of Medicine, aims to facilitate the use of biomedical terms and concepts in computerized information systems. UMLS offers several software tools, including the Metathesaurus, a large, multilingual vocabulary database that crosswalks and categorizes codes and concepts from other classification systems and code sets. The structural and semantic properties of UMLS are robust enough to use for exploring relationships among different concepts (Patel and Cimino 2009).

Licensees of the UMLS Metathesaurus have access to the Systematized Nomenclature of Medicine Clinical Terms, generally known as SNOMED CT. Owned and maintained by the International Health Terminology Standards Development Organisation in Denmark, SNOMED CT is a multilingual terminology developed to retrieve and code clinical information from EHRs reliably (Elkin et al. 2006; Cornet and de Keizer 2008; Lee, Lau, and Quan 2010). This nomenclature is part of a suite of US government–designated standards for the electronic exchange of clinical health information. Other widely used classification schemes include those developed by the World Health Organization (Chapter 5) and groups interested in specific scientific areas (e.g., laboratory testing, genetics). Many of these classification schemes' codes, concepts, and terms do not overlap and must be bridged using methods such as the UMLS Metathesaurus (Pathak et al. 2009). For some clinical areas, classification approaches are less developed. As consensus builds over the next several years on the best methods for coding clinical data from EHRs and other HIT sources, ensuring that the classification systems thoroughly reflect the full range of health conditions will be important.

ICUs are already largely automated—they gather electronic information from physiological monitoring devices and other digital sources—and therefore are leading other clinical settings in electronic computation of risk scores. Junger and colleagues (2002) examined the Sequential Organ Failure Assessment (SOFA), a completely automated risk score calculation developed by the European Society of Intensive Care Medicine. The computerized SOFA captured laboratory values relatively easily, but level of consciousness, a critically important factor in assessing ICU risk, was more difficult to extract from electronic data sources. Because information needed to compute Glasgow Coma Scores is often embedded in free-text physical examination notes, Junger and colleagues (2002) designed structured query language

(SQL) scripts to retrieve relevant information from physicians' clinical examination texts. The SQL scripts also helped identify artifacts from the automated vital signs monitor. For instance, during routine flushing of arterial lines, the automated monitor falsely recorded zero values for blood pressures, which would seriously distort SOFA scores.

Natural language processing techniques are another option for extracting clinical information from narrative texts (Wang et al. 2008). Researchers apply natural language processing to EHRs to identify clinical scenarios of interest, such as use of particular medications (Wang et al. 2009; Pacheco et al. 2009). Automated data mining and natural language processing could also help make problem lists in EHRs more complete (Chen et al. 2008). An automated problem list created from free-text EHRs using natural language processing may also improve the efficiency and accuracy of diagnosis coding (Meystre and Haug 2006). Furthermore, natural language processing techniques applied to EHRs could provide timely information about brewing epidemic illnesses (Hripcsak et al. 2009) or medication complications (Wang et al. 2009). Whether natural language processing could identify clinical indicators for risk adjustment remains unknown.

Patients' Roles in EHR Documentation

The role of patients in documenting their health histories in EHRs is evolving, but it will undoubtedly increase in coming years. Studies show that patients are willing to provide information about their health in electronic formats. In the mid-1990s, Wald and colleagues (1995) designed a health history interview that queried patients about physical and psychiatric symptoms, health-related behaviors, and various risk factors. New patients were asked to come early for scheduled appointments to answer interview questions at computer terminals in the waiting room. Patients' answers were downloaded directly to their EHRs. The computerized interview took an average of 27 minutes. Although 42 percent of older patients had no prior keyboard experience, overall patients rated the computerized interview positively; 65 percent preferred the computer-administered survey, whereas 15 percent preferred face-to-face interviews (Wald et al. 1995, 149). Patients willingly answered sensitive questions: 13 percent reported serious domestic violence in the prior 12 months, and 16 percent reported suicidal ideation.

A more recent study used the Internet to gather data from patients. The researchers developed a series of 232 primary questions (to be asked of all patients) and 6,000 additional questions (to be asked only of patients chosen on the basis of their responses to prior questions) (Slack et al. 2011). They tested the reliability of the primary questions on a sample of patients to see whether patients provided identical responses over time. They found adequate

test-retest reliability, suggesting that Internet-based history questionnaires could be useful to patients for documenting their clinical histories in EHRs.

An increasing number of providers are offering Internet portals protected by encryption software that enable patients to communicate with their clinicians by secure e-mail, examine test results, look at medication and problem lists, conduct transactions (e.g., schedule appointments, order prescription refills), and read their EHRs. Patient examination of EHRs is a sea change in health care. Many concerns about this new practice remain, but evidence of its benefits is building (Delbanco et al. 2010, 123):

> Descriptive studies and trials of patient-accessible medical records have begun to evaluate the effect of patient access to notes, but most allow patients to review paper medical records only in a clinical setting. Overall, the literature suggests that patient access to the medical record may improve patient–doctor communication, empower and educate patients, and foster adherence. Risks seem minimal, and few patients find their records confusing or anxiety-provoking, in part perhaps reflecting self-selection. Patients generally report few concerns about documenting sensitive issues (for example, sexuality, marital problems, and drug use), and providers report that patients may correct some serious inaccuracies. Physician concerns about effects on documentation, their own time, and staff time are common, but in practice, no substantial effects have been demonstrated.

How patients will respond to increased access to their EHRs is another question that warrants further study. Delbanco and other researchers (2010) in Boston spoke to a convenience sample of patients in advance of an experiment giving them increased access to their EHRs. The researchers received a complex mix of responses (Delbanco et al. 2010, 123):

> Patients also voiced pros and cons. Some clearly did not want to read what their doctors wrote because they were worried about discovering something they would rather not know, finding potential diagnoses that might make them anxious, or reading what their doctors really thought of them. Others feared reading something that would shake their trust in their doctors. Some felt that unfamiliar medical terminology would make them misinterpret what they read. They wondered how to learn to ask the right questions and who should teach them. More than a few noted that their doctors are already stretched to the limit, and if they were required to write notes that patients can read and understand, time for examination and consultation could diminish. Some also worried that ready access would compromise privacy if electronic information ended up in the wrong hands.
>
> However, our patients also anticipated benefits. As more patients e-mail their doctors and use other online services, some saw open communication through electronic notes as a logical next step. Others thought that reviewing, and ultimately contributing to, visit notes could help them get on the same page as their doctors. Many expected to search for explanations of technical language on the Internet. Some believed their

> doctor's notes would prove educational simply by reminding them of what happened during the visit. They expected some notes to reassure them and to calm their fears; other notes might be "truth tellers" and push them to face the reality of a health issue, such as obesity and mental illness, and perhaps break down defenses. Many liked the idea of sharing notes with family, friends, partners, and informal consultants, anticipating that this would help build a personal care system at home.

Regardless of how the current generation of patients feels about access to their EHRs, it is inevitable that the next generation of patients, with their inborn facility with electronic information systems in general, will feel comfortable with electronic health information. The development of personal health records—electronic health information that patients can carry wherever they go—could also have important implications for medical record documentation (Fricton and Davies 2008; Dimick 2010; Krist and Woolf 2011). In short, patient involvement will likely alter, in as yet uncertain ways, the content of medical records. The utility of this new breed of patient-filtered health information in risk adjustment is a question for future researchers to explore.

Notes

1. This study looked at deaths among patients hospitalized for acute myocardial infarction, congestive heart failure, stroke, gastrointestinal hemorrhage, or pneumonia and among patients who underwent abdominal aortic aneurysm repair, coronary artery bypass grafting surgery, or craniotomy (Pine et al. 2007).
2. The researchers developed mortality prediction models for stroke, pneumonia, myocardial infarction, heart failure, and septicemia (Tabak, Johannes, and Silber 2007).
3. The 39 conditions for which risk-adjusted mortality models were developed were ischemic stroke, hemorrhagic stroke, asthma, chronic obstructive pulmonary disease, tuberculosis, pneumonia, aspiration pneumonia, respiratory failure, pulmonary embolism, arrhythmia, arterial aneurysm, venous disorder, valvular heart disease, acute myocardial infarction, chest pain, heart failure, catastrophic vascular disease, catastrophic cardiovascular disease, other circulatory disease, esophageal disease, gastroduodenal disorder, intestinal disease, gastroenterologic bleeding, gastroenterologic obstruction, gastroenterologic perforation, infectious diarrhea, colorectal cancer, other gastroenterologic disease, gallbladder disease, acute liver disease, liver/bacillary disease, pancreas disease, hip/upper femur fracture, diabetes, chronic renal disease, acute renal failure, urinary infection, sepsis, and high-risk infection (Tabak et al. 2010).

CHAPTER

7

DATA FROM PATIENT SURVEYS

Lisa I. Iezzoni and Karen Donelan

In proposing fundamental redesign of the US health care system, the Institute of Medicine (IOM 2001a) viewed patients' perspectives as the "true north" guiding radical health care change (Berwick 2002). Among IOM's six suggested aims for health system reform is patient centeredness: "providing care that is respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions" (IOM 2001a, 6). Section 6301 of the 2010 Patient Protection and Affordable Care Act (ACA; see Chapter 1) adopted this focus in creating the nonprofit Patient-Centered Outcomes Research Institute (PCORI) to "assist patients, clinicians, purchasers, and policy-makers in making informed health decisions." The law sets priorities for PCORI's research, directing that gaps in evidence related to "patient needs, outcomes and preferences" be addressed. Furthermore, PCORI "shall provide support and resources to help patient and consumer representatives effectively participate" in its governance and decision making. As the ACA recognizes, the only way to ensure and evaluate patients' preferences is to ask patients for feedback.

Providers, health plans, employers, and governmental agencies ask people lots of questions about their health and health care. Some worry that the flurry of surveys is overwhelming a public weary of intrusive telemarketing and spam. One Massachusetts health plan abandoned its annual survey of enrollees' satisfaction with their physicians when the response rate fell to 10 percent (Safran and Rogers 2002, 9). Declining response rates reflect not only a weary public but also changing technology. In 2010, 29.7 percent of all US households had wireless telephones only, a reality that has dramatically changed the cost and quality of survey sampling and survey research (Blumberg and Luke 2011).

Nonetheless, the complexities of collecting health care quality measures from other data sources (administrative information or medical records) make obtaining information directly from patients seem relatively simple. One convenience is the availability of accurate lists of patients and their current contact information, both of which health care providers and payers generally have. As described in chapters 5 and 6, other data sources cannot provide complete or accurate information about patients' health and health care experiences. For this reason, performance measurement initiatives frequently focus

on patients' reports about their health care experiences or their health or functional performance. Like comparison of other outcome measures, meaningful comparison of these findings across providers generally requires risk adjustment. Assessment of patients' perceptions about their care experiences typically is based on risk factors from patient-reported information, sometimes supplemented with administrative data sources (e.g., see O'Malley et al. 2005; Eselius et al. 2008; Johnson, Rodriguez, and Solorio 2010; Palsbo et al. 2010). Patient-reported data about functional status are also used to risk-adjust certain Medicare capitation payments (Kautter, Ingber, and Pope 2008).

This chapter explores issues concerning the use of data from patient surveys for risk adjustment. It does not review in-depth technical attributes of survey methodologies, sampling procedures, sample sizes, or analytic approaches but instead focuses on aspects pertinent to risk adjustment. It draws examples primarily from efforts to risk-adjust capitation payments using functional status information and the Consumer Assessment of Healthcare Providers and Systems (CAHPS). Government health and health care surveys, such as those shown in Exhibit 7.1, are frequently used for outcomes research. Results are risk adjusted by data elements reported by survey respondents.

EXHIBIT 7.1 Examples of Health and Health Care Surveys

	Organization	Population	Main Topics	Website
Current Population Survey (CPS)	US Census	National, state	Employment, income, and health insurance	www.census.gov/cps
Behavioral Risk Factor Surveillance System (BRFSS)	CDC	National, state, city, county	Health risk behaviors, preventive health practices, health care access, and chronic disease and injury	www.cdc.gov/BRFSS
Youth Risk Behavior Surveillance System (YRBSS)	CDC	National, state, school district	Health risk behaviors, including tobacco and alcohol use, sexual activity, diet and exercise, and injury and violence	www.cdc.gov/HealthyYouth/yrbs/index.htm
National Immunization Survey (NIS)	CDC	National, state, local estimates	Child immunization records and household and provider data	www.cdc.gov/nis
National Health Interview Survey (NHIS)	CDC, NCHS	National	Health insurance, immunization, use of health services, and behavioral and genetic risk factors	www.cdc.gov/nchs/nhis.htm

EXHIBIT 7.1 (Continued)

	Organization	Population	Main Topics	Website
National Health Care Surveys (NHCS)	CDC	National	Collection of surveys on hospital and ambulatory care	www.cdc.gov/nchs/nhcs.htm
National Long Term Care Survey (NLTCS)	NIA and Duke University	National	Longitudinal survey on health and functional status of persons aged 65 or over living in communities and institutions, service use, and caregiver resources	www.nltcs.aas.duke.edu
National Health and Nutrition Examination Survey (NHANES)	NCHS	National	Patient interview and physical examination data on the health and nutritional status of children and adults	www.cdc.gov/nchs/nhanes.htm
Consumer Assessment of Healthcare Providers and Systems (CAHPS)	AHRQ	National	Patient experiences with ambulatory care and hospitalization	www.cahps.ahrq.gov/default.asp
Medical Expenditure Panel Survey (MEPS)	AHRQ	National	Health care costs/expenditures, access to care, health disparities, healthcare quality, and child and elder health	www.meps.ahrq.gov/mepsweb
Medicare Current Beneficiary Survey (MCBS)	CMS	National	Medicare beneficiary experience with coverage and health services	www.cms.gov/MCBS
Health Information National Trends Survey (HINTS)	NCI	National	Public search for and use of health information	http://hints.cancer.gov
Survey Data Archives				
Roper Center Public Opinion Archives	Roper Center, University of Connecticut	Multiple	Large data repository for health research	www.ropercenter.uconn.edu/data_access/data/data_providers.html
Inter-University Consortium for Political and Social Research (ICPSR)	University of Michigan	Multiple	Large data repository for social science surveys	www.icpsr.umich.edu/icpsrweb/ICPSR

AHRQ: Agency for Healthcare Research and Quality
CDC: Centers for Disease Control and Prevention
CMS: Centers for Medicare & Medicaid Services
NCHS: National Center for Health Statistics (within CDC)
NCI: National Cancer Institute
NIA: National Institute on Aging

Importance of Risk-Adjusting Patient-Reported Outcomes

With certain exceptions (e.g., the functional status information used to risk-adjust capitation payment levels), the outcomes assessed through patient health surveys typically include self-reports of health status, such as disease, disability, and physical function, and perceived outcomes of care, such as experiences of care in physicians' offices and hospitals. Thus, patients generally define the "what" in the "risk of what?" question (see Chapter 2). Survey-based initiatives can consider more diverse outcomes than can administrative data–based projects, in which outcome measures are derived from items in fixed data sets, such as mortality, costs, and service use. As described later in this chapter, the major constraints facing survey-based initiatives are appropriate design; logistical feasibility (including cost); and respondents' availability, physical ability, and willingness to answer specific questions.

Some patient-reported outcome instruments are generic (i.e., relevant to all or most persons regardless of health conditions), whereas others are condition or situation specific. Therefore, when aiming to assess patient-reported outcomes, analysts first select the eligible population and the process or outcome measures of interest and then identify appropriate instruments and data collection methods for gathering outcomes information of the highest possible quality. Numerous tools are available for soliciting patient-reported outcomes, such as overall health status, quality of life, symptoms, satisfaction with care, and adherence to therapy. Launched in 2002, the Patient-Reported Outcomes Measurement Information System (PROMIS; see Chapter 3) is a network funded by the National Institutes of Health that is compiling sets of dynamic tools to capture patient-reported functional status and other outcomes (www.nihpromis.org/default.aspx). According to the PROMIS website, the goal of this initiative is to build and validate "common, accessible item banks to measure key symptoms and health concepts applicable to a range of chronic conditions." To organize its extensive item banks, PROMIS groups its measures into domains and uses item response theory and other psychometric techniques to analyze items and scales (Cella et al. 2007; Hays et al. 2007; Liu et al. 2010). The MAPI Institute (www.mapi-institute.com), headquartered in Lyon, France, translates patient-reported outcome measures into various languages, striving to ensure conceptual equivalency. According to its website, the MAPI Institute has translated and "linguistically validated" more than 500 instruments into over 150 languages. Validation is context specific, something potential users should consider when selecting instruments.

The need to risk-adjust patient-reported outcomes depends on the context. Risk adjustment of comparisons across groups of patients may mask differences associated with important patient characteristics (see Chapter 3). However, failure to risk-adjust could bias these comparisons. Research suggests

that patients' health and health care experiences are affected by many factors, including their age, sex, race and ethnicity, education, health literacy, burden of illness, extent of functional impairments, cultural and religious beliefs and practices, and expectations and preferences for health care—potentially most of the risk factors described in Chapter 3. The extent to which patient factors contribute to patients' perceptions of care varies by context. Sometimes patient factors are poor predictors of outcomes (e.g., see Davies and Cleary 2005; Kressin et al. 2010; Kong et al. 2007). Safran and Rogers (2002, 24) argued, "The research literature shows that measured patient characteristics explain little of the variation in patients' assessments of care—most often 5 percent or less." Nonetheless, others urge risk-adjusting for patient factors, especially when the goal is to influence clinician behavior. Clinicians are likely to believe the data only after they are risk-adjusted for factors they associate with case complexity (Hong et al. 2010).

Risk adjustment of patient-reported outcomes raises sensitivities because patients from historically disadvantaged subgroups have been more likely to report poor outcomes. Race and ethnicity have attracted particular attention due to well-documented disparities in care (IOM Committee on Understanding and Eliminating Racial Disparities in Health Care 2002; Rodriguez et al. 2008). Studies have shown that nonwhite patients are often less satisfied with their care and experience worse health outcomes than do white patients, although the extent and direction of these differences varies by racial and ethnic subgroup and the outcome of interest.

Efforts to characterize the role of black race in patient-reported outcomes have produced provocative findings. LaVeist, Nickerson, and Bowie (2000) oversaw telephone interviews in which white and black respondents with cardiac disease were asked about their perceptions of racism in the health care system, trust of medical providers, and satisfaction with care. Black patients were significantly less satisfied with their care than were white patients, but once perceptions of racism and mistrust of medical providers were added to multivariable regression models, black race no longer significantly predicted dissatisfaction. Also using telephone survey data from white, black, and Latino respondents, Chen and colleagues (2005) found that black and Latino individuals who strongly perceived racial discrimination in health care were significantly more likely to prefer racially and ethnically concordant physicians. Black persons who preferred racially concordant physicians and had a black physician were significantly more likely than black patients with nonblack physicians to rate those physicians as excellent. Among Latino individuals, the association between patients' racial concordance with their physicians and rating their physicians as excellent was not statistically significant, although the data trended in that direction. The results of this study contrasted with the findings of a study of black and white ischemic heart disease patients at five VA medical centers that examined eight scales summarizing health-related beliefs, attitudes, and experiences (Kressin et

al. 2002),[1] including whether patients felt respected by physicians and treated with dignity. Only one scale—attitudes toward religion—yielded statistically significant differences between black and white respondents: Black patients "placed stronger importance on God and religion in general, as well as in coping with their heart problems and in making decisions about their heart disease" (Kressin et al. 2002, 179).

Controlling for race and ethnicity can therefore affect perceptions of patients' experiences with care. Farley and colleagues (2011) tried to understand why California had much lower CAHPS scores for both fee-for-service and Medicare managed care plans than did the rest of the country. Although findings had been adjusted for the standard CAHPS case-mix variables, these variables do not include race and ethnicity. Addition of race and ethnicity and urbanicity indicators improved the relatively lower ratings of California plans. A study of patient experiences across 33 community health centers in Washington state found that adjusting for race and ethnicity and patients' language altered rankings of the centers (Johnson, Rodriguez, and Solorio 2010).

Women and men may have different expectations for care. Even after controlling for gender-based expectation differences, one study found that women reported lower satisfaction levels than did men (Mitchell and Schlesinger 2005).

Self-perceived health and functional status are often critical risk factors for predicting patient-reported outcomes. Studies have found that persons who have poor health (Hargraves et al. 2001; Schlesinger, Druss, and Thomas 1999; Keenan et al. 2009) or disabilities (Iezzoni et al. 2002; Palsbo et al. 2010) are generally less satisfied with their medical care. According to Iezzoni and colleagues (2002), "one [potential] explanation is that people with greater needs for services have more interactions with the health care system and therefore more opportunities to be disappointed. Certainly, people with significant health problems are more likely than others to need timely care involving a range of services and to want information about their conditions, prognoses, and therapeutic options. With more treatment decisions and interventions comes more chance of mishap involving technical and interpersonal quality of care."

In certain subpopulations, important risk factors for satisfaction with care are context specific. One study involving 2,652 veterans found that membership in veterans' organizations, use of non-VA services, and military combat exposure were significantly associated with satisfaction (Harada, Villa, and Andersen 2002). Other important risk factors included race and ethnicity and functional status. Not surprisingly, dissatisfied veterans are more likely to change their sites of care, potentially disrupting care continuity—a particular problem for veterans with chronic illnesses (Stroupe et al. 2005). In another study, researchers used responses from approximately 4,000 patients enrolled in behavioral managed care plans to risk-adjust the findings from their patient

experience survey (Eselius et al. 2008). The final risk adjustment model included self-reported mental health and general health status, alcohol and drug treatment, age, education, race, and ethnicity. Although risk adjustment only modestly affected scores of the 21 behavioral health plans studied, it did change some plan rankings.

As noted in Chapter 3, patient preferences and expectations for health care are potentially important predictors of such outcomes as satisfaction with care. Strongly expressed preferences sometimes have little effect on the care patients receive. For example, one multicenter study found that patients' desires did not determine their place of death: Even patients who strongly preferred to die out of hospital had a similar likelihood of dying in hospital to that of others who had no preference or did not express their preferences as strongly (Pritchard et al. 1998). Desires or expectations for specific elements of care, such as laboratory tests, physical examinations, and referrals to specialists, may be unrelated to sociodemographic characteristics (Zemencuk et al. 1999). However, the concordance between such desires and expectations and what patients actually receive may strongly influence satisfaction with care.

Regardless of the statistical association between patient-reported outcomes and risk factors, the choice of appropriate risk factors, if any, depends on the context. Commenting on assessment of physician performance through patient surveys, Safran and Rogers (2002, 24) raise fundamental questions about risk-adjusting physician-specific results:

> A dilemma that we face in adjusting survey-based performance data for patient characteristics is that, in some ways, it runs counter to efforts to promote patient-centered care. That is, by leveling the playing field with adjustment, we negate the differing needs of patients served by different physicians, rather than holding physicians accountable for how well they serve their particular patient mix. Nevertheless, it is almost certainly imperative that we risk-adjust physician-level performance data. Without this, the data will be readily disregarded by those who feel that their results are biased downwards by the particular characteristics of their patient panel.

Collecting Risk Factor Information Through Surveys

As noted in the previous paragraphs, many critical risk factors for patient-reported outcomes are historically sensitive topics, such as race and ethnicity, or represent concepts such as functioning and overall health, which seem straightforward but are often interrelated in complex ways with gender, culture, personal preferences, and other factors. Surveys that reflect and represent the experiences of populations at higher risk for adverse health status and outcomes with some statistical reliability and validity are challenging to conduct.

Appropriate survey methods can capture the experiences of persons with a higher burden of illness due to disability, poor health, poverty, literacy, or communication disorders, but it may be costly to obtain data that are timely enough to be actionable from these populations. Survey design, measurement, and data collection methods must aim to ensure that the experiences of all patients are represented without bias.

Conducting surveys that produce credible measures of patients' experiences and their risk factors therefore requires time, expertise, and resources. Nevertheless, for various reasons, many surveys are not designed or executed to represent fully the entire patient population. Because of cost, convenience, and other factors, many surveys exclude people who are difficult to survey due to age, language, disability, or other reasons. Furthermore, low response rates may compromise the integrity of survey results. Proper use of scientific survey sampling and data collection methods ensures that representative random or probability samples reflect the experiences of a population with a measurable degree of accuracy and error.

Several outstanding texts are available to inform the development of high-quality surveys; in addition, guidance is available through professional associations of survey researchers, including the American Association of Public Opinion Research (AAPOR), the American Statistical Association (ASA), and the Council on American Survey Research Organizations (CASRO). The Office of Management and Budget (OMB), which approves surveys conducted for the federal government, also provides useful summaries of issues regarding the development, design, testing, conduct, and analysis of surveys. The considerations described in the following sections are especially critical to assessing the quality of survey data.

Questionnaire Development and Measurement

Individuals willingly respond to many questions about outcomes (e.g., health status, perception of experiences) and risk factors. Developing a questionnaire *de novo* requires considerable expertise, time, testing, and resources. Ideally, questions should be derived by asking persons from the target population detailed questions about the outcome of interest. Draft questions should be tested to ensure that respondents understand the questions as the researchers intended. The draft questionnaire also requires pilot testing, including cognitive testing and possibly data analyses (e.g., factor analysis, Chronbach's alpha), to explore interrelationships among different questions or survey responses. Draft questionnaires may undergo multiple revisions and refinements before they are finalized.

Because of the complexities of designing new questionnaires, use of an existing instrument that suits the purpose is an efficient option. As noted earlier and in Chapter 3, numerous survey questionnaires and scales that capture the health experiences of patients (e.g., instruments accessed through

the PROMIS website) are available for use. PROMIS provides available evidence on the statistical reliability and validity of its posted measures. However, when assessing measures developed for use in other studies, researchers need to examine the population for which that measure was developed and tested. For instance, surveys validated in English or in geographic regions where few immigrants or ethnic minorities live may not be useful for studying more diverse populations. Scales used for persons without functional impairments may not apply to persons with disabilities. At times, the imperative of reaching diverse populations may require that questionnaires be developed and tested on different samples of respondents or using different modes of administration than originally used. In these circumstances, qualitative research to develop and test key concepts, terms, and constructs may be important, and pretesting is essential (Donelan 2009).

Sometimes a person's health status is a barrier to reliable reporting (e.g., a survey respondent might have cognitive impairment or disordered memory). In these situations, more objective assessments may be necessary. The proximity of the time of the survey to the period of recall for person-reported measures is important. For persons at most ages, memories fade with time, although certain sentinel events (e.g., the birth or death of a family member, a new major diagnosis) may be recalled for long periods. Ideally, clinical details that must be gathered with specificity (dates, procedures, medication names) should be taken from clinical records or at least validated against the records of a sample of patients. For subjective questions about feelings, perceptions, or experiences, accuracy is neither measurable nor necessary; the respondent's answers represent "truth" at that place and time.

For respondents who are not able to answer questions on their own for reasons of age, cognition, disability, or language, some researchers include the responses of proxies. Proxies are frequently used in surveys of children, depending on the nature of the questions. Proxy respondents are not always able to report on issues such as patient attitudes, but nonetheless, appropriate proxy respondents who have been closely involved in the care of a family member or friend may provide valuable information on decision making, the utilization of services, and other factors. Family members and caregivers are often chosen as respondents in surveys about end-of-life care (e.g., see Parsons, Baum, and Johnson 2000; Hendershot 2003; Henry J. Kaiser Family Foundation 2003).

When language is a barrier to patient participation, surveys must never be "sight-translated" by interpreters or family members, as this approach will compromise the standardization that is critical to survey measurement. One should not assume that a patient has an appropriate level of literacy to read and respond to self-administered survey questions in his native or primary language. The assistance of certified translation experts who have experience in health or medical settings is critical; online translation services or

other tools may render interpretations that are inappropriate in health and medicine.

Capturing Key Risk Factors

As noted earlier, some risk factors for patient-reported outcomes are sensitive attributes or complex concepts, and careful attention must be paid to operational definitions. As suggested in Chapter 3, capturing key risk factors through surveys can be complex. For example, concerns about gender identity (e.g., transgender individuals) can complicate straightforward questions about respondents' sex. Another example is OMB Statistical Policy Directive No. 15, which sets federal standards for gathering racial and ethnic data (see Chapter 3); its requirement that separate questions be asked about race and Hispanic or Latino ethnicity could confuse respondents. While OMB standards are commonly used, they are not universal. A review of federal surveys used for health care and epidemiological research found considerable inconsistency in reporting of race, ethnicity, and immigration status (Johnson, Blewett, and Davern 2010). Likewise, definitions of language proficiency, socioeconomic status, insurance coverage, and other factors differ across existing questionnaires, and these differences can greatly affect survey findings and the implications of results. Therefore, understanding the exact meaning and implications of all questions is essential before selecting a survey measure "off the shelf."

Representativeness

Survey scientists must first consider the population of interest when designing methods to produce data that represent it. Designers of high-quality surveys target a well-defined population, select respondents from a database/source that maximizes coverage of that population without introducing bias, and sample the population using methods that yield a known or equal probability of selection of different individuals in that population (Donelan 2009). For institutional or practice-specific surveys of patient populations, designers can use lists from records or administrative systems maintained by those patients' providers, recognizing that certain patients might get some of their care from providers other than the particular provider from which the patients are sampled. Community-based surveys aim to produce findings that represent a given community; they might therefore collect data from people where they live and work. But randomly drawing respondents to represent communities is a complex task. Some survey designs systematically exclude or include certain segments of the population.

The mode of recruiting or surveying subjects could potentially produce a biased sample. For example, telephone surveys may systematically exclude people with hearing impairments but may be more appropriate for people with vision loss. An increasing proportion of the population, especially

young adults, has only cell phones, so they would be excluded from traditional random-digit-dialed landline telephone surveys. Surveys conducted over the Internet exclude individuals without computer access, including disproportionate numbers of elderly, lower-education, and low-literacy subgroups. Nonetheless, Internet-based surveys may be the best way to reach young adults. Mail surveys are inexpensive and easy to use for sample populations that have good contact information (i.e., reliable and up-to-date addresses) but are inappropriate for persons with poor literacy, persons who are periodically away from home, or homeless individuals.

Thus, a variety of factors affect the representativeness or generalizability of a survey and the quality of data a survey produces. The central question is whether the survey is accessible to all members of the population of interest; if it is not, characteristics of those likely to be excluded must be determined. Practical questions include whether the lists of sample respondents include accurate contact information and, if not, whether those missing this information represent a particular subgroup (i.e., information is not missing at random). Finally, does the mode of conducting the survey systematically exclude certain subgroups in the population of interest?

Sample Size and Sample Design

High-quality surveys target sample sizes sufficient to minimize sampling error and ensure sufficient statistical power for analyzing the outcome measures. For surveys to adequately represent the experience and views of persons at higher risk for health problems, complex sampling designs, including the use of disproportionate sampling techniques, are often necessary. Simple random samples generally do not produce reliable estimates of these populations, unless they are very large—and thus very costly (Donelan 2009).

For example, suppose a hospital decides to survey 500 patients about their care experience and is especially interested in the views of its elderly patients. If 10 percent of the hospitals' patients are elderly and about 50 percent of sampled patients complete the survey, only 25 elderly individuals will likely respond. The number could be smaller if for some reason (e.g., mode of survey administration) elderly individuals are less likely to complete the survey. Analyses based on 25 elderly patients would produce too much error to provide representative data. If instead the hospital drew a sample of 250 elderly and 250 non-elderly patients, it would use similar resources but produce more reliable information about elderly patients' experiences.

Thus, choice of sampling design depends on the answer to several questions, including whether any population subgroups are of particular interest. More complex sampling strategies than strict random sampling may be necessary to obtain sufficient numbers of responses for meaningful analyses involving subgroups of interest.

Response Rates

As noted at the outset of this chapter, surveys currently suffer from declining response rates. Desirable response rates from populations at higher risk for disease and disability have always been more challenging to achieve. Many individuals in this population are especially hard to reach, regardless of mode of survey administration (e.g., mail, telephone, Internet, in-person household surveys) (Donelan 2009). Low-income individuals, including persons lacking health insurance, may not open their mail or answer the telephone (fearing bill collectors) or may be unable to pay telephone or Internet charges. Home foreclosures and inability to pay rent may cause persons to move, making their addresses hard to track. Diligent and costly efforts may therefore be required to reach individuals whose life circumstances are chaotic—a subgroup that might be of special interest with regard to health problems. Nonfluency in English, disability, and other factors may also make people difficult to reach.

To achieve good response rates, survey efforts may require multiple and varied respondent contacts, several modes of administration, prolonged data collection periods, strategies for converting nonrespondents and refusals into respondents, and appropriate financial or other incentives. Researchers may calculate and report response rates in different ways. Many medical and health care research journals require a minimum response rate to make survey reports acceptable for publication. Many surveys conducted today do not meet these standards but are still used frequently by organizations, the media, advocacy groups, and others.

An extensive literature on response rates shows that lower response rates do not necessarily mean that estimates from surveys are biased (Keeter et al. 2006; Groves 2006). Nonetheless, data from surveys to which less than 50 percent of potential respondents reply should be interpreted cautiously. Major peer-reviewed journals generally view properly calculated response rates higher than 60 percent more favorably. Professional organizations such as the American Association for Public Opinion Research (www.aapor.org) have expended considerable effort to encourage standard reporting of response rate calculations and to indicate strategies that induce responses most effectively and lend credence to data.

Cost

The costs of collecting risk factor information from surveys must be weighed against the expense of using other methods for gathering these data. Katz and colleagues (1996) were concerned about the logistics and expense of paying trained nurse abstractors to gather medical record information for the Charlson Comorbidity Index (see discussion later in the chapter). They therefore developed a questionnaire to obtain information directly from patients by mail or through an interview. The relative costs of obtaining comorbidity information by patient report versus chart review were estimated as follows:

$0.93 for mailed, self-report questionnaires; $1.67 for patient interviews; and $3.50 for chart abstraction (Katz et al. 1996, 79). These comparative cost estimates are certainly out of date, and electronic health records may expedite collection of clinical information from medical records (see Chapter 6). When exploring the feasibility of gathering data through surveys, researchers need to seek cost estimates from different survey vendors and compare what each offer covers.

Studies of Accuracy of Patient-Reported Risk Factor Data

For many years, patients' voices were rarely included in analyses of health systems and health services (Donelan 2009). Some questioned whether patients fully observe or understand what happens in health care settings and whether they can judge the quality of health care services. In fact, patients can supply information about almost all potential risk factors, except perhaps physiological values and technical clinical findings. Surveys gathering patient-reported outcomes therefore generally include questions about potential risk factors, even ones as broad as overall health. Some patient surveys are linked to other data sources, such as health insurance enrollment or claims files, that generate the sampling frame and contain information about possible risk factors (e.g., coded diagnoses). In certain circumstances, confidentiality concerns demand that surveys be completely anonymous (i.e., precluding linkage to other sources). In these instances, survey responses provide the only insight into risk factors.

Patient-reported risk factors fall along a continuum from putatively objective or measurable attributes, such as the presence of diseases or health behaviors (e.g., smoking), to those that are intrinsically subjective, such as self-perceived health status. For subjective measures, assessing accuracy is moot: Their very subjectivity, a synthesis of individual views and preferences, is their value. Despite the subjectivity of measures such as global self-perceived health, however, they can significantly predict important outcomes ranging from mortality to annual expenditures (Bierman et al. 1999). However, according to Cleary (1997, 4):

> Some researchers are uncomfortable with subjective variables because they are perceived as unreliable. Such people often think of data from medical records as "hard" data, whereas they think of survey responses as "soft" data. Thus, rather than judging the relative theoretical value of objective and subjective measures, some researchers' selection of variables is unduly influenced by their negative opinions about the value of survey data relative to other types of information. However, medical records contain many types of data, including information about subjective states collected using unstandardized methods.

Efforts to examine the accuracy of patient-reported risk factors have generally focused on more "objective" items, typically comparing patients' reports to a medical record gold standard. Comparisons find that patients are most accurate when asked about well-defined medical conditions (e.g., diabetes, heart attacks) and least accurate about less specific conditions (e.g., arthritis) (Silliman and Lash 1999). In general, older patients and those less educated provide less accurate reports than do younger, well-educated patients. Older patients' reports may be incomplete, especially when they have coexisting conditions (Gross et al. 1996).

One concern is the possibility that nonrespondent bias affects the generalizability of specific health information obtained from surveys. Primatesta and colleagues' (2007) guide to conducting population-based surveys about cardiovascular disease and its risk factors warns about the need for high response rates because the health characteristics of nonrespondents tend to differ from the health characteristics of those answering such surveys.

Studies looking at the accuracy of patient-reported risk factors have considered a range of clinical areas. One study compared patients' reports of cardiovascular disease risk factors obtained through a telephone interview to the findings gathered by physicians and nurses in a subsequent clinic visit (Bowlin et al. 1996). Reported risk factors were compared to objective test results; for example, they measured the patients' exhaled carbon monoxide levels to identify smokers, measured the patients' blood pressure three times, and measured serum total cholesterol. On the basis of clinical gold standards, approximately half of the respondents with hypertension and hypercholesterolemia misclassified themselves as not having the risk factor, while reports of smoking were generally accurate. In another study, Yeager and Krosnick (2010) analyzed data from the National Health and Nutrition Examination Survey comparing participants' reported use of nicotine-containing products to serum cotinine levels (an indicator of nicotine exposure) and found that few of the participants (around 0.9 percent) underreported nicotine use. In the RAND Medical Outcomes Study, physicians' reports of "tracer" diagnoses (e.g., diabetes mellitus) agreed more with confirmatory laboratory data (e.g., glycosylated hemoglobin) than did patients' reports in self-administered questionnaires or face-to-face interviews (Kravitz et al. 1993). A study of VA clinic patients compared self-reported chronic diseases to diagnoses documented in medical records. Patients accurately reported 97 percent of chronic obstructive pulmonary disease and 95 percent of coronary artery disease (Fan et al. 2002).

One study from the Netherlands found that persons with mobility limitations reported conditions that were not noted by their primary care physicians, such as chronic lung and cardiac disease, atherosclerosis, cancer, stroke, and arthritis (Kriegsman et al. 1996, 1411), raising concerns about overreporting of health conditions. A study that used the Canadian National

Population Health Survey, which prospectively follows a cohort of respondents, found that only 41 percent of participants accurately reported episodes of major depression that had occurred ten years earlier, while 90 percent reported episodes that had occurred in the prior year (Patten et al. 2011). This study suggests a potential for underreporting some mental health conditions.

Some studies have compared numerical risk adjustment scale results derived from patient reports with those based on medical record reviews. Katz and colleagues (1996) developed a questionnaire containing the Charlson comorbidity factors (see Chapter 3) that participants could self-administer or answer in an interview. Kappa statistics (see Chapter 9) indicating agreement between patients' reports of individual comorbidities and medical record evidence of comorbidities ranged from 0.35 to 0.85. Charlson scores from patient-generated data produced a Spearman correlation coefficient of 0.63 when compared to scores derived from medical records (Katz et al. 1996, 77). In a study of women with early-stage breast cancer, Silliman and Lash (1999) found high correlations (roughly 0.6; $p < 0.001$) between comorbidity scores derived from telephone interviews and those derived from medical record reviews. Nevertheless, scores based on patient interviews were generally slightly higher than those from medical record review, suggesting that the interviews identified more comorbid conditions than did chart abstraction. Another study used patient self-reports of diagnoses to produce a Charlson Comorbidity Index and used diagnoses from administrative data to derive Charlson scores for the same patients (Chaudhry, Jin, and Meltzer 2005). Although both indexes predicted one-year mortality well, the administrative data-based model performed slightly better.

Other researchers have examined the accuracy of persons' reports of processes of care or health care events. For example, drug use may indicate the severity or complexity of diseases such as diabetes. One study compared self-reports of care among persons with diabetes to medical record and claims information (Fowles et al. 1999). In a subanalysis of drug use, the investigators found that self-reported insulin administration was virtually 100 percent accurate; about 10 percent of respondents overreported and 10 percent underreported use of oral hypoglycemic agents.

In contrast, although hospitalization seems a memorable event, comparisons of patients' reports and administrative records show that patients systematically underreport hospital admissions (Clark, Ricketts, and McHugo 1996; Roberts et al. 1996; Wallihan, Stump, and Callahan 1999; Ritter et al. 2001). Not surprisingly, persons' memories are especially faulty after considerable time elapses. Individuals are likely to underreport use of physician services, emergency room use, and hospitalization. One study of 422 adults aged 60 or older found that 24 percent failed to report a hospitalization they had within the prior year, and 28 percent did not report their visit to an emergency room (Wallihan, Stump, and Callahan 1999, 662).

Accuracy of recall of cancer screening services has generated considerable research. Rauscher and colleagues (2008) conducted a meta-analysis of the validity of self-reported cancer screening histories and found that sensitivity was the highest for mammography, clinical breast examination, and Pap smears and the lowest for prostate-specific antigen testing and digital rectal exams. Summary estimates of sensitivity and specificity were lower among predominantly black and Hispanic samples than among predominantly white samples. "When estimates of self-report accuracy from this meta-analysis were applied to cancer-screening prevalence estimates from the National Health Interview Survey, results suggested that prevalence estimates are artificially increased and disparities in prevalence are artificially decreased by inaccurate self-reports" (Rauscher et al. 2008, 748).

For certain types of risk factors, patient-reported information may be more accurate and easier to obtain than information from clinicians or medical records. Surveys are especially valuable for capturing concepts such as functional status (Cleary 1997, 3–4):

> Patients can be asked whether they have difficulty going up and down stairs, or an observer can visit their homes to observe whether they can or cannot climb stairs. This is a situation in which objective measures are available and can be more reliable and valid, if properly administered, than patient self-reports, but such methods are often prohibitively expensive. . . . Medical records frequently contain functional assessments that were obtained by health care professionals with no training in standardized measurement and that are largely subjective measures. Such variables may be measured more efficiently, reliably and validly with standardized subjective measures.

Patients may also be the best source of potentially sensitive information. Once patients report this information, it also becomes uniformly available through other sources (e.g., medical records). For example, a pilot test of the Partners in Health Survey, which involved randomly telephoning households in central Pennsylvania, found missing values of only 0.9 percent for smoking, 1.0 percent for consuming at least one alcoholic drink in the last month, and 2.3 percent for keeping a loaded and unlocked firearm (Bazos et al. 2001). However, 21.6 percent of income data were missing.

Issues Relating to Surveys of Population Subgroups

Many studies are limited to respondents who speak English, given the costs and linguistic challenges of translating questionnaires (i.e., ensuring that the translation has the same meaning as the English version). Rigorous techniques for translating questionnaires require multiple translations from English to other languages and back again. Cross-cultural differences in

attitudes toward health, symptoms, and disease are crucial considerations in translating surveys for use across nations or across linguistic groups within populations.

Cultural differences may affect response categories, such as the words used to weight either extreme of a Likert-type scale. For instance, the global, self-perceived health status question is typically phrased: "How would you rate your health? Excellent, very good, good, fair, or poor?" During a study of Chinese immigrants who spoke Mandarin or Cantonese, Ngo-Metzger and colleagues (2003) found that respondents perceived "excellent" as an almost superhuman quality virtually unattainable by average persons. They changed the word weighting the upper end of the scale to one meaning roughly "very, very, very good." Cultural differences may also affect graphic representations of survey concepts. King and colleagues (2011) aimed to design a questionnaire about acute coronary syndrome for administration to European (white), South Asian, and Chinese patients. Assessments by culturally concordant lay reviewers determined that the graphic used to depict squeezing pain differed for different groups: European and South Asian participants responded to a hand grasping a balloon, while Chinese participants preferred the image of wringing out a towel.

Perceptions of health conditions may also vary across cultures. For instance, one project looked at self-reports of global visual functioning by cataract patients in the United States, Manitoba (Canada), Barcelona (Spain), and Denmark. The participants were asked to rate the trouble they were experiencing with their eyesight as "a great deal," "moderate," "a little," or "none" (Alonso et al. 1998, 870). The researchers measured actual visual acuity using standard Snellen-type charts. Among respondents with 20/40 or better visual acuity, 34.5 percent of US respondents reported "a great deal" of trouble with vision, as did 32.1 percent of those from Denmark; in contrast, only 22.2 percent and 14.8 percent of persons from Barcelona and Manitoba, respectively, responded this way (Alonso et al. 1998, 873). Across all respondents, comorbidity significantly increased the risk of reporting trouble with vision.

Designing survey methodologies for persons with disabling conditions requires special thought. Beyond concerns about sampling persons with disabilities and considering the appropriateness of proxy respondents, significant logistical questions arise (Parsons, Baum, and Johnson 2000; Mitchell et al. 2006). Interview and telephone surveys of persons who are deaf or hard of hearing are especially challenging. Telephone surveys could employ teletypewriter (TTY) or Telecommunications Device for the Deaf (TDD) equipment, but these technologies pose serious impediments to the survey process. For instance, sometimes the length of typed text allowed is too short to accommodate entire questions. In addition, persons whose primary language is American Sign Language may not have complete facility communicating in

English, similarly to other linguistic minorities. New video technologies and relay services could facilitate surveys of persons requiring hearing accommodations. The design of survey materials for persons who are blind or have low vision must consider alternative presentation modalities, such as Braille, large print, or audio. Persons with speech impairments who use communication technologies also require special accommodations.

Surveys administered through traditional landline telephone services may fail to reach a disproportionate number of young adults (persons aged 18 to 30) and certain racial and ethnic minorities who increasingly use only cell phones (Frankel et al. 2003; Lee et al. 2010; Call et al. 2011). Although using weighting and other adjustments might allow some surveys to account for differences relating to cell phone–only households (Frankel et al. 2003), these adjustments may not account fully for differences relating to young adults (Call et al. 2011). "While careful weighting can mitigate noncoverage bias in landline telephone surveys, the rapid growth of cell-phone population[s] and their distinctive characteristics suggest it is important to include a cell-phone sample. Moreover, the threat of noncoverage bias in telephone health survey estimates could mislead policy makers with possibly serious consequences for their ability to address important health policy issues" (Lee et al. 2010, 1121).

Using Survey Data for Risk Adjustment in Health Policy Settings

Published studies about patient-reported outcomes are growing. Most such studies look at associations between outcomes and potential risk factors or risk-adjust the outcomes for comparison across subgroups of patients. Survey-based risk adjustment has become important in health policy contexts, notably for determining appropriate capitation payments and evaluating plan performance. Surveys are especially suited to the managed care environment, where respondent populations are well circumscribed and contact information (e.g., names, addresses, telephone numbers) is typically available. Two such initiatives are reviewed briefly in the following sections.

Using Frailty Adjustments to Set Capitation Payments

Interest in using survey-based data to risk-adjust capitation payments to managed care organizations (MCOs) arose from fundamental concerns about the limited clinical content of the administrative data typically used for this purpose. Survey data could presumably provide essential information missing from administrative data, especially for the most severely debilitated subgroup of enrollees. The 1997 Balanced Budget Act required Medicare to risk-adjust reimbursement for chronically ill and disabled enrollees to pay for

them more fairly, thereby linking MCOs' capitation payments to health status. However, ICD-9-CM diagnosis codes do not adequately represent functional deficits, such as those experienced by frail elderly persons and disabled Medicare beneficiaries. Payments may not reflect the true costs of caring for particularly vulnerable persons if all aspects of patients' clinical presentations are not considered. If MCOs do not receive adequate payments, they may avoid these needy patients.

Medicare already uses functional status information to pay certain providers (see chapters 15 and 16). Medicare's prospective payment systems for nursing homes, home care agencies, and inpatient rehabilitation facilities use functional status information provided by clinicians rather than information obtained directly from patients. In these care settings, clinicians routinely record patients' functional status. Therefore, although gathering these data for payment purposes could be administratively burdensome, the necessary documentation about functional status generally already exists.

Researchers have examined whether functional status and other patient-reported information, alone or combined with data on diagnoses, improves predictions of costs for persons enrolled in MCOs. Results have been mixed. Some studies have found that self-reported functional status information gathered from SF-36 questionnaires or the Medicare Current Beneficiary Survey produces substantially better predictions than do models using only demographic or diagnostic information (Hornbrook and Goodman 1995, 1996; Gruenberg, Kaganova, and Hornbrook 1996). However, other studies have found that risk adjustment using diagnoses, specifically ACGs and DCG/HCC models, produces better predictions than does self-reported health status information (Fowles et al. 1996; Pope et al. 1998; Iezzoni and Greenberg 2003).

The frailty adjustment for Medicare's Program of All-Inclusive Care of the Elderly (PACE) is one model that uses survey data to determine capitated reimbursement (Kautter, Ingber, and Pope 2008; Hirth, Baskins, and Dever-Bumba 2009). Authorized by the Balanced Budget Act of 1997, PACE provides comprehensive services financed through integrated Medicare and Medicaid capitation. It was modeled on a system of acute and long-term care services pioneered by On Lok Senior Health Services in San Francisco and tested through demonstration programs in the 1980s. To be eligible for PACE, Medicare beneficiaries must be at least 55 years old, live in a PACE service area, and be certified as eligible for nursing home care by their appropriate state agency (CMS 1998).

Research found that the demographic and diagnosis-based risk adjustment method used to set capitation rates for Medicare MCOs—the HCC model—underpredicted costs for PACE members with activity of daily living (ADL) limitations. Therefore, Medicare now requires PACE programs to survey their members using the Health Outcomes Survey-Modified (HOS-M)

to gather information for use as a frailty adjuster to set capitated payment rates (CMS 2010c, 114):

> The HOS-M instrument is a shorter, modified version of the Medicare Health Outcomes Survey and contains 6 ADL items as the core items used to calculate the frailty adjustment factor for payment purposes. The survey also includes 12 physical and mental health status questions from the VR-12 [Veterans RAND 12-Item Health Survey]. In addition, the HOS-M includes questions about the following: lifting or carrying objects as heavy as 10 pounds; walking a quarter mile; health or physical problems interfering with daily activities, receiving help with ADLs; physical and emotional health compared to one year ago; memory loss; urinary incontinence; and a question on whether the survey was self-completed or completed by a proxy. If the participant received assistance completing the survey, the respondent was asked information about the proxy respondent.

HOS-M data facilitate the annual adjustment of plan payment rates for the frailty of each PACE organization's enrollee panels (CMS 2010c, 113).[2] Capitated payments from Medicare are set as follows (CMS 2010c, 131):

> The risk score is computed for each participant for a given year and applied prospectively. The risk score generally follows the beneficiary for one calendar year. The prior year's functional impairment data are used to predict the next year's payment adjustment. Functional data are submitted to CMS, where they are calculated to establish the PACE organization's frailty score, which is then applied to each participant's risk adjusted payment. The frailty adjustment approach is applied in conjunction with the CMS–HCC risk adjustment model.

The capitated Medicaid payment to each PACE program is negotiated separately by each state. It also takes into account the frailty of PACE enrollees that has been established through the use of survey data.

Consumer Assessment of Healthcare Providers and Systems (CAHPS)

CAHPS is one of the largest-scale efforts ever launched to gather patient-reported outcomes information. The Agency for Healthcare Research and Quality (AHRQ) launched this now three-phase program in the mid-1990s with a consortium of researchers to focus on gathering consumer and patient experiences with health care plans, institutions, and providers. CAHPS surveys emphasize patients' experience of care using standardized measures designed to ensure valid comparisons across health care settings. Sound scientific principles, input from all affected parties (patients, providers, payers), and public transparency are key guiding principles of the process, consistent with the best survey practices (see discussion earlier in the chapter).

Details about CAHPS survey instruments, the National CAHPS Benchmarking Database, uses of CAHPS for quality improvement, and other topics are available on the AHRQ website (www.cahps.ahrq.gov). At the time of this

writing, CAHPS offers the following survey instruments (new questionnaires are periodically released):

- The **CAHPS Health Plan Survey** asks enrollees about their recent experiences with services and their health plans. Versions are available for adults and children covered by commercial insurers, Medicare, or Medicaid (see Hargraves, Hays, and Cleary 2003; Kim, Zaslavsky, and Cleary 2005).
- The **CAHPS Clinician & Group Survey** asks persons about their recent experiences with physicians and physicians' staffs. Versions are available for adults receiving primary care or specialty services and for children receiving primary care (see Rodriguez et al. 2009; Gallagher et al. 2009).
- The **CAHPS Surgical Care Survey**, developed by the American College of Surgeons and the Surgical Quality Alliance, asks patients about their experiences with surgeons, surgeons' staffs, anesthesiologists, and surgical care before, during, and after surgery (see Hoy 2009).
- The **Experience of Care and Health Outcomes (ECHO) Survey** asks adult health plan enrollees about their experiences with behavioral health care and services provided by managed behavioral health care plans or MCOs (see Beebe et al. 2003).
- The **CAHPS Dental Plan Survey**, developed for the TRICARE dental plan, asks adult enrollees in dental care plans about their experiences with their plans, dentists, and dentists' staffs (see Keller et al. 2009).
- The **CAHPS American Indian Survey**, developed for Choctaw Nation Health Services, asks adult American Indians about their experiences with their health care (see Weidmer-Ocampo et al. 2009).
- The **CAHPS Hospital Survey** asks adult patients about recent experiences with inpatient medical, surgical, or obstetrical care (see Goldstein et al. 2005; Giordano et al. 2010).
- The **CAHPS Nursing Home Survey** includes three separate questionnaires: (1) an in-person structured interview with long-term residents, (2) a mail survey for recently discharged short-stay residents, and (3) a mail survey for family members of residents (see Sangl et al. 2007).
- The **CAHPS Home Health Care Survey** asks adults receiving home care services about their experiences with home care.
- The **CAHPS In-Center Hemodialysis Survey** asks adults with end-stage renal disease about their experiences with the facility where they receive dialysis.

The following supplemental item sets can be appended to their corresponding survey:

- The **CAHPS Item Set for Children with Chronic Conditions** (for use with the children's version of the CAHPS Health Plan Survey) asks

parents or guardians to report on the health care experiences of children with special health care needs.

- The **CAHPS Item Set for People with Mobility Impairments** (for use with the adult version of the CAHPS Health Plan Survey) asks about the health care experiences of adults with impaired mobility.
- The **CAHPS Item Set for Addressing Health Literacy** (for use with the CAHPS Clinician & Group Survey) asks adults to report on their clinicians' ability to communicate health information.
- The **CAHPS Health Information Technology Item Set** (for use with the CAHPS Clinician & Group Survey) asks patients about their health information technology experiences relating to physician care.
- The **CAHPS Cultural Competence Item Set** (for use with the CAHPS Clinician & Group Survey) asks patients about the cultural competence of their clinicians.

One of AHRQ's goals is to use the data from the CAHPS surveys to provide consumers useful information to help them choose among health plans. Risk adjustment of CAHPS results is therefore key to producing valid comparisons across health plans or providers. As noted earlier in this chapter, the topics assessed by CAHPS surveys—perceptions of various health care experiences—could be strongly associated with respondents' characteristics, such as their age, gender, and education. CAHPS collects some, albeit limited, information that can serve as risk factors. For its early version, "[t]he general CAHPS recommendation, based on a literature review and analyses from CAHPS demonstration sites, is to adjust for age and health status" (Zaslavsky et al. 2000, 166). CAHPS asks respondents to rate their health status on a five-point scale from excellent to poor. The CAHPS risk adjustment recommendation was drawn from studies showing that younger persons and those in worse health provide more negative evaluations of their care than do older and healthier persons (Iezzoni et al. 2004). CAHPS researchers found that Medicare MCOs differed significantly in the overall health status of their enrollees (Zaslavsky and Buntin 2002). Nevertheless, early work showed that adjusting for respondents' risk factors (in some models including educational status and proxy respondent) had little effect on comparative CAHPS performance across health plans (Zaslavsky, Zaborski, and Cleary 2000). Another study controlled for age, health status, education, and interactions with specific health plans; it found that risk adjustment decreased the variability of ratings across plans but did not dramatically alter perceptions of plan performance (Elliott et al. 2001).

In examining early CAHPS results, researchers found that respondent bias could skew the findings. CAHPS health plan surveys among Medicare beneficiaries achieved overall response rates of 75 to 80 percent, much higher than for many other such surveys (Zaslavsky, Zaborski, and Cleary 2002).

However, older beneficiaries, persons with disabilities, women, racial and ethnic minorities, and persons living in geographical areas with relatively high numbers of poorly educated and impoverished persons had lower response rates to mailed surveys. Inaccurate contact information explained some of these differences. Conducting telephone interviews of persons who did not return mailed surveys improved the representativeness of respondents. For-profit health plans were significantly more likely than nonprofit plans to have inaccurate contact information and lower response rates. Most worrisome, "Plans with lower ratings on the CAHPS survey also had more bad contact information" and consequently lower response rates (Zaslavsky, Zaborski, and Cleary 2002, 497). "If experiences with health care affect propensity to respond, nonresponse might bias plan comparisons. The fact that CAHPS scores from respondents are related to plan response rates suggests that this might be happening" (Zaslavsky, Zaborski, and Cleary 2002, 497).

At the time of this writing, information from the AHRQ CAHPS website indicates that benchmarked scores are adjusted for several "case mix" factors. For the CAHPS Clinician & Group Survey, the risk adjustment model includes respondent age, education, and self-reported health status. Although CAHPS analysts considered adjusting for mode of survey administration, data are not yet available to support adjustment for this factor for the Clinician & Group Survey. Using linear regression, the CAHPS Health Plan Survey ratings (but not the item-level data and frequencies) also adjust for case mix by accounting for age, education, and self-reported health status.

Aside from these three case-mix factors, research found that other patient characteristics affect CAHPS ratings, including race and ethnicity (Weech-Maldonado et al. 2001, 2003; Goldstein et al. 2010; Farley et al. 2011); respondents' language (Weech-Maldonado et al. 2003, 2008; Setodji et al. 2011); disability (O'Day et al. 2002; Palsbo et al. 2010); and geographic location (Weech-Maldonado et al. 2008; Mittler et al. 2010; Keenan et al. 2010; Farley et al. 2011). However, controlling for a risk factor could mask its effects on comparisons of performance across plans or providers. Because CAHPS restricts its risk adjustment to a parsimonious three factors, observers need to interpret comparisons across plans or providers cautiously; many factors in addition to differences in quality performance could explain observed differences in CAHPS ratings.

Notes

1. The eight scales addressed self-perceived disease severity, patient evaluation of physician, evaluation of VA care, attitudes toward religion, satisfaction with decision making, perceived urgency of catheterization, vulnerability to catheterization, and bodily impact of catheterization.

2. For the HOS-M, a random sample of Medicare beneficiaries is drawn annually from each participating PACE plan that has at least 1,400 members and surveyed each spring (CMS 2010c, 113). PACE members are considered eligible for the HOS-M if they reside in the community, do not have end-stage renal disease, and are aged 55 or older. All eligible members of PACE organizations with panels fewer than 1,400 are surveyed.

CHAPTER

8

CONCEPTUAL AND PRACTICAL ISSUES IN DEVELOPING RISK ADJUSTMENT METHODS

Lisa I. Iezzoni

Development of credible risk adjustment methods is challenging, costly, and time consuming. It is therefore easiest to take an existing method "off the shelf" if it suits the purpose. Even if existing methods do not perfectly match a project's goals and context, the trade-off may be worth it. The convenience of an existing risk adjustment method may outweigh the costs of developing and validating a new approach. When a method does not precisely fit a project's setting and purpose, analysts must interpret results cautiously to recognize conceptual mismatches and the potential need for statistical recalibration. Nonetheless, the availability of existing methods, even imperfect ones, is powerfully attractive, especially for producing timely results.

Sometimes, however, existing methods are inadequate. As noted in Chapter 1, most widely used risk adjustment methods aim to meet broad policy objectives, such as setting payment levels or comparing hospital mortality or complication rates. Often, the audiences for which the risk-adjusted results are intended are willing to tolerate certain limitations, such as minimal clinical detail. When the target audience is clinicians, however, clinical credibility is vitally important. Use of clinically valid, transparent risk adjustment methods is especially crucial when trying to motivate clinicians' behavior (e.g., to improve quality or change practice patterns) or convince them of the merit of particular therapies (e.g., using data only from observational studies rather than from randomized controlled trials). Furthermore, in narrow or specific clinical settings, generic risk adjustment methods intended for use across broad populations might fail to capture critical complexities. In these instances, development of a new risk adjustment method may be necessary.

This chapter describes important considerations in developing risk adjustment methods. Again, I emphasize the complexity of this undertaking and urge potential users to consider existing methods. In doing so, however, it is important to recognize the statistical derivation of most risk adjustment methods today (as opposed to measures developed more than two decades ago, which frequently relied on normative clinical judgment to set weights for individual risk factors). Given these statistical origins, one must ensure

that the risk adjustment method is appropriately calibrated to the population or setting of interest. For example, the Agency for Healthcare Research and Quality (AHRQ) Quality Indicators (QI) use a reference population to produce their risk adjustment parameters (AHRQ 2010a).[1] AHRQ's QI software uses indirect standardization and applies smoothing algorithms after estimating O/E ratios (ratios of observed-to-expected complication rates) to generate risk-adjusted, "smoothed" rates. Several situations could disrupt the applicability (or calibration) of AHRQ's QI algorithms to different settings, including the following circumstances (AHRQ 2010a):

- The data are more recent than those from AHRQ's reference population, and adverse outcome rates have consistently improved over time. In this case, O/E ratios would average less than 1.0 when models based on the older reference population are applied to newer populations.
- The data come from hospitals not included in AHRQ's reference population (e.g., Veterans Affairs [VA] medical centers, hospitals in states not represented in AHRQ's database), and adverse outcome rates at these hospitals are consistently better or worse than outcomes at hospitals in AHRQ's reference population. In this case, O/E ratios would average more or less than 1.0.

In both circumstances, systematic recalibration of the risk-adjusted, smoothed rates is generally indicated.

This chapter addresses diverse issues, including specifying predictor variables, the role of clinical judgment, and empirical modeling techniques. It does not contain the detailed technical discussions (e.g., on multivariable modeling) available in statistical textbooks. Instead, this chapter focuses on issues pertinent to creating risk adjusters. To make this discussion real, I use examples from two older but exceptionally well-done projects to develop clinically based risk adjustment methods. Although these studies occurred up to 25 years ago, they hold lessons for today. The earlier project developed the Medicare Mortality Predictor System (MMPS), a medical record–based method for risk-adjusting Medicare beneficiaries' hospital mortality rates for four conditions: acute myocardial infarction (AMI), congestive heart failure (CHF), pneumonia, and stroke (Daley et al. 1988).

The larger second study developed risk adjusters to identify adverse surgical outcomes for patients in VA hospitals. This project began in 1991 as the National Veterans Affairs Surgical Risk Study in 44 VA medical centers (Khuri et al. 1995, 1997; Daley et al. 1997a, 1997b) and by 1993 encompassed 87,078 major noncardiac operations. The investigators used multivariable logistic regression to empirically derive risk adjusters predicting all-cause mortality and specific morbidities (complications) within 30 days after the index procedure for all operations and eight surgical subspecialties. In 1994,

the risk adjustment techniques derived in the Surgical Risk Study became part of the National Surgical Quality Improvement Program (NSQIP), an ongoing initiative of quality assessment activities in the VA (Khuri et al. 1998; Daley, Henderson, and Khuri 2001). NSQIP now resides with the American College of Surgeons, and it is a major source of new insights about risk adjustment methods and the value of risk-adjusted outcomes in surgical quality improvement (Hall et al. 2009; Dimick et al. 2010; Ingraham et al. 2010a, 2010b; Raval et al. 2010; Turner et al. 2011).

Getting Started

The first step in considering risk adjustment involves answering four fundamental questions (see Chapter 2): Risk of what outcome? Over what time frame? For what population? For what purpose?

One fundamental reality is that research projects or measurement initiatives can study only outcomes for which data are available. Historically, this need has limited their focus to mortality, costs (or charges), use of specific procedures or services, complications or adverse events, and patient-reported satisfaction with care—a modest subset of the full range of potentially important health care outcomes. Little if any information is systematically available about patients' signs and symptoms, functional status, and quality of life or whether care met patients' stated or implicit goals. This situation may change with increasing use of electronic health records (EHRs), especially as patients themselves begin entering information into these files (see Chapter 6). Already data collection for some NSQIP variables is largely electronic: Laboratory data are automatically transmitted from participating hospitals' electronic laboratory systems to the central NSQIP database.

Outcomes must be clearly and reliably defined. Reliability—and sometimes the ability to audit data to ensure its integrity—is critically important when outcomes data are used for public reporting or determining payment levels (e.g., in a pay-for-performance scheme). For example, in addition to tracking postoperative mortality, NSQIP considers multiple morbidity outcomes, including superficial and deep wound infections, postoperative pneumonia, unplanned intubation, pulmonary embolism, acute renal failure, stroke, myocardial infarction, sepsis, coma, and bleeding more than four units of blood. NSQIP developers spent considerable effort defining each of these outcomes so that data would be gathered reliably and consistently across institutions. Early on, some outcomes eluded clear, consistent definition and were therefore jettisoned (e.g., postoperative ileus, peripheral neurological injury).

From their clinical experiences, the NSQIP investigators expected each morbidity outcome to generate its own set of risk factors. Even similar outcomes—for example, postoperative respiratory failure and postoperative

pneumonia—yielded slightly different sets of risk adjusters from empirical modeling with NSQIP data. Important risk factors for respiratory failure included type of surgery, age, whether surgery was performed emergently, low albumin, high blood urea nitrogen (BUN), dependent functional status, and history of chronic obstructive pulmonary disease (Arozullah et al. 2000, 250). Pneumonia generated similar risk factors, but with notable variations (e.g., different ranges for BUN, surgical types), and additional factors, such as alcohol intake and blood transfusions (Arozullah et al. 2001).

The outcome must also be frequent enough to be usable for statistical modeling. Even in a massive database such as the NSQIP file, some outcomes can be too rare to support separate risk adjustment modeling. A good example was modeling postoperative mortality for transurethral prostatectomy and other urological procedures. Although NSQIP developmental data contained more than 200,000 of these procedures, mortality was so low that stable models were impossible to produce. Numbers plummeted further when the performance of individual hospitals was examined, the primary goal of NSQIP. To deal with infrequent individual outcomes, researchers often try various approaches to aggregating different outcomes into a single super-outcome or composite measure. For instance, early NSQIP investigators considered several ways to combine 21 morbidity indicators, including the total number of morbidities reported, weights derived from regressing postoperative length of stay on the 21 morbidities, and mean ratings by 44 chiefs of surgery on a scale of 1 to 5 of the likelihood that the morbidity could cause death, disability, prolonged hospitalization, or patient dissatisfaction (Daley et al. 1997b; Gibbs et al. 1999). This simple summary measure—the presence or absence of one or more of the 21 morbidities—performed just as well in predicting postoperative length of stay as did the more complicated metrics.

Optimal assessment of the effectiveness of care requires investigation of positive and negative outcomes. Positive outcomes include amelioration of symptoms, improved functional status, prevention of death, and lower costs. Adverse outcomes include mortality, complications, lack of improvement or worsening functional status, dissatisfaction with care, and higher costs. Trade-offs among these outcomes may arise, especially between clinical benefits and resource use. In some scenarios, better clinical outcomes (e.g., survival, functional ability) are possible but at a high price, such as lengthy stays in neonatal intensive care units among low-birth-weight newborns.

Because gathering data on risk factors is expensive, even in an increasingly digital environment an important question is whether a minimum or core set of risk factors can be used to adjust for risks of different outcomes, such as mortality, complications, and costs. Considerable research has assessed this hypothesis. Although important risk factors are often similar across different outcomes, relationships between these factors and particular outcomes (i.e., "weights," empirical parameter estimates) vary by outcome. In one study, my

colleagues and I used forward stepwise regression techniques to select the ten most statistically important MedisGroups key clinical findings (KCFs) to predict hospitalization costs and in-hospital death among patients at least age 65 (Iezzoni, Moskowitz, and Ash 1988). For the three conditions studied (stroke, pneumonia, and AMI), the ten most important risk factors varied for predicting cost versus death (Exhibit 8.1).

The high cost of gathering risk factor and other data is a major obstacle to attracting more hospitals to participate in NSQIP. In the 2005–2007 database, which included information from 241 participating centers (Dimick et al. 2010, 504):

> More than 130 patient and operative variables are recorded, including patient demographics, preoperative risk factors, patient laboratory values, intraoperative variables, and postoperative 30-day morbidity and mortality. The data collection process relies on a sampling strategy aimed at collecting a diverse set of operations. Trained surgical clinical nurse reviewers record the data using standardized definitions. The reliability of the data is ensured through intensive training mechanisms for the surgical clinical nurse reviewers and by conducting inter-rater reliability audits of participating sites.

EXHIBIT 8.1 Top-Ten Key Clinical Findings Using Stepwise Regression Models to Predict Costs and In-Hospital Death for Three Conditions

	Stroke		Pneumonia		AMI	
Step	Cost	Death	Cost	Death	Cost	Death
1	Oxygenation	Coma	Oxygenation	BUN	CHF	Cardiac arrest
2	Lethargy	Respiratory rate	Hematocrit	Cardiac arrest	BUN	BUN
3	Creatinine	Cardiac arrest	Calcium	Systolic blood pressure	Cardiac arrest	Oxygenation
4	Myocardial ischemia	Stupor	Carbon dioxide	Arterial pH	Creatinine phosphokinase	Systolic blood pressure
5	Edema	Glucose	Calcium squared	Stupor	Positive sputum culture	Coma
6	Potassium	Pulse	Cardiac arrest	Respiratory rate	Cardiomegaly	Bundle branch block on ECG
7	Wheezing	BUN	Lethargy	Coma	Calcium	AMI on ECG
8	Partial thrombo-plastin time	High-density brain mass	Premature ventricular contractions	Alkaline phosphatase	Potassium	Stupor
9	Calcium	Systolic blood pressure	Sodium	Pleural effusion	Pulse	Atrial flutter
10	AMI on ECG	History of diabetes	Blood in stool	Arterial pH squared	AMI on ECG	Positive sputum culture

Note: For physiological parameters that can have both high and low abnormal values, the value of the parameter squared as well as the raw value were entered into the model.

To improve the efficiency and focus of NSQIP, plans are in place to move from sampling all procedures in a specialty to gathering information on 100 percent of a small number of core procedures. This change would reduce the range of variables reviewers need to collect. Furthermore, NSQIP analysts explored using a parsimonious set of variables rather than the complete set of variables (i.e., variables with p value < 0.1, up to 70 variables) in the risk adjustment models (Dimick et al. 2010, 504; see Exhibit 8.2 for the top five variables):

> We created an intermediate variable set by combining the 5 most important variables, in terms of order of entry in the stepwise regression model, for each procedure. Because of substantial overlap in the variables included for each procedure, there were only 12 unique variables for mortality and 11 unique variables for morbidity. . . . To determine the

EXHIBIT 8.2 Predictor Variables from Stepwise Logistic Regression Models: National Surgical Quality Improvement Program Data, 2005–2007

Order of Importance	Predictor Variables by Surgical Procedure				
	Cholecystectomy	Ventral Hernia Repair	Gastric Bypass	Pancreatectomy	Colectomy
Mortality models					
1	Functional status	Functional status	Functional status	Functional status	ASA class
2	ASA class	Dialysis	CHF	CHF	Functional status
3	Ascites	ASA class	Hypertension	ASA class	Emergency
4	Weight loss	Wound class	Gender	Dyspnea	Albumin
5	Albumin	Dyspnea	Dyspnea	Albumin	Dyspnea
Morbidity models					
1	ASA class	ASA class	Functional status	Functional status	ASA class
2	Functional status	Functional status	Bleeding disorder	ASA class	Functional status
3	Wound class	Wound class	ASA class	Bleeding disorder	Emergency
4	Ascites	Dyspnea	Diabetes	Dyspnea	Albumin
5	Albumin	BMI	CHF	BMI	BMI

ASA = American Society of Anesthesiologists
BMI = body mass index
Source: Adapted from Dimick et al. (2010, 505).

impact of even more limited risk-adjustment models on risk adjustment, we created a limited model using the 2 most important variables from each procedure-specific model, which included only 5 variables (ie, American Society of Anesthesiologists [ASA] class, functional status, congestive heart failure, dialysis, and bleeding disorder).

In terms of predictive power and discrimination, the NSQIP models using five and even two predictor variables performed similarly to the models based on the full set of clinical variables. For instance, for mortality predictions for cholecystectomy, the c-statistics of the full, intermediate, and limited models were 0.93, 0.93, and 0.92, respectively. For mortality predictions for gastric bypass, the c-statistics of the full, intermediate, and limited models were 0.84, 0.83, and 0.79, respectively (Dimick et al. 2010, 506). Although these NSQIP analyses suggest an exceedingly parsimonious model performs as well as a complete model, the models' performance may differ in other clinical contexts. A few highly influential factors may drive patients' risks for surgical mortality and complication outcomes, but the same may not be true for other patient populations and outcomes.

The characteristics of the study population have implications for the risk adjustment strategy (see Chapter 3). Research targeting one particular disease (e.g., diabetes) does not need to adjust for the presence of the disease, although adjustment for disease-specific severity (e.g., extent of end organ damage, such as retinopathy, peripheral vascular disease, or renal insufficiency) may be important. Skewed sex distribution within the population could also raise concerns, a frequent consideration in studies of VA populations because, until recently, the majority of veterans were men. For example, in the VA Surgical Risk Study, 96.7 percent of the 87,000+ surgical patients were male (Daley et al. 1997b, 331). However, because of the large sample size, the number of women (3.3 percent of the population) was sufficiently large, so retaining them in the analytical database did not cause statistical problems. As study populations expand (e.g., comparisons across states or regions), adjustment for risk factors that are likely to be distributed evenly across large populations (Gatsonis et al. 1995) becomes less necessary. However, the numbers of patients admitted to individual hospitals are too small to ensure random distribution of risk factors; one usually should assume that patients admitted to different hospitals vary in a nonrandom fashion, making risk adjustment necessary for comparing their outcomes. Even populations admitted to large academic medical centers are not sufficiently large for comparison of their outcomes to eschew risk adjustment.

Whether risk adjustment should be condition specific or generic (diagnosis independent) depends on the context. In the NSQIP, risk adjustment models differed even for such closely related complication outcomes as postoperative respiratory failure and pneumonia. With increasing sophistication in risk adjustment methods and improvement in data quality, the recent trend

has been toward developing disease-specific risk models. The evolution of APACHE over three decades is illustrative. Developed to predict the probability of in-hospital death for ICU patients, the original APACHE computed a generic score combining information about age, chronic diseases, and acute physiological parameters (Knaus et al. 1981). Because no ICU data set existed at the time, development of this initial APACHE relied largely on clinical judgment. Ten years later, based on a database containing thousands of cases, the empirically derived APACHE III calculated scores using the same algorithm for any patient, regardless of diagnosis, admitted to an ICU. However, this algorithm for predicting probability of in-hospital death considered whether patients fell into 78 disease groups (Knaus et al. 1991). In APACHE IV, 116 diagnostic categories (compiled from the full list of 430 admission diagnoses) are used in the algorithm (Zimmerman et al. 2006).

Finally, often the ultimate aim of risk-adjusted outcome analyses is to evaluate quality of care and motivate improvement. This motivation has important implications for risk adjustment and the overall analytic strategy. Designers of clinically credible, statistically robust risk adjustment methods must recognize that the units of observation, such as hospitals, physician practices, and individual physicians (e.g., cardiothoracic surgeons), may vary widely in size and patient mixes. For outcomes information to prompt behavior change, all health care professionals who must contribute to improvement initiatives must feel comfortable that the risk adjustment method is clinically meaningful.

Identifying Risk Factors

Sometimes choices of risk factors are limited. For example, administrative data offer few options. Gathering additional information is generally infeasible unless one links the database to another administrative source (see Chapter 5). Many risk adjustment methods based on administrative data exist, however, so taking one off the shelf may be more appropriate than developing an entirely new method. Even if an existing method does not exactly suit the purpose, components of these methods could prove useful. For instance, as described in Chapter 3, several existing algorithms group ICD-9-CM diagnosis codes into clinically meaningful categories. Empirical weights are then assigned to these categories through multivariable modeling. Two major examples of this strategy are groupings developed by Elixhauser and colleagues (1998) from AHRQ and the Hierarchical Condition Categories (HCCs) (Pope et al. 2004). Elixhauser and collaborators identified 30 conditions associated with length of stay, hospital charges, and in-hospital deaths (see Exhibit 3.7). Their method is freely available on the AHRQ website (www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp). Developers of the HCCs sorted

inpatient and outpatient diagnosis codes into 189 condition categories and then used them to predict current and future one-year costs. The full HCC model is proprietary, but the trimmed-down CMS-HCC model employing a selected set of HCCs, used to set payment for capitated Medicare plans, is freely available on the CMS website (www.cms.gov/MedicareAdvtgSpecRateStats/06_Risk_adjustment.asp). Investigators can enter ICD-9-CM grouped variables into their own models, allowing each condition variable to find its own weight in their data sets (see Chapter 3).

Predictive modeling tools—statistical techniques used to predict health outcomes, sometimes called "health care forecasting" techniques—are becoming increasingly powerful and popular (Forrest et al. 2009). Even with administrative data, however, it is essential to develop *a priori* clinical hypotheses about how potential risk factors relate to the outcome(s) of interest. Otherwise, empirical modeling to develop a risk adjuster could lapse into "data dredging"—allowing the computer to identify key risk factors and specify their relationship to the outcome (i.e., the coefficients or weights). Although data dredging can yield models with compelling *p*-values and strong performance metrics (e.g., R^2 values, c-statistics) in the developmental data set, these models may not validate well in other data sets (see chapters 9 and 10) and may lack clinical credibility (i.e., they may suggest relationships between risk factors and outcomes that do not make clinical sense). If clinicians do not believe in the risk adjustment method, they are unlikely to trust the results.

Researchers can use several strategies to identify candidate risk factors and develop hypotheses about how each relates to the outcome(s) of interest. Published reports from randomized trials and other clinical studies may help. The clinical literature, however, is richer in some fields than in others. In particular, randomized controlled trials (RCTs) are relatively limited, forcing reliance on observational studies. In any case, RCTs generally provide little information about risk factors for predicting costs, functional status, and quality of life.

Another strategy is to ask clinical experts or panels of practicing clinicians. Clinical knowledge derived from patient care experience, as well as experts' syntheses of the literature, is critical to informing model development and ensuring its ultimate acceptance. Clinician involvement in developing risk adjusters promotes clinical credibility. Individual interviews with clinical experts may also generate hypotheses about important risk factors. In developing the MMPS in the mid-1980s, Daley and colleagues (1988) convened subspecialty physician panels and provided them with complete literature reviews concerning patient risk factors for 30-day mortality. They identified panelists on the basis of recommendations from subspecialty societies and other major research projects concerning the four target diseases. Leaders of the VA Surgical Risk Study convened expert panels to identify patient risk factors for postoperative mortality and morbidity and to group similar surgical procedures according to

likelihood of adverse outcomes. Chiefs of surgery at the participating hospitals became deeply involved in overseeing all aspects of risk adjustment and other methodological decisions, which led the chiefs and other surgeons to accept the study results.

The "herding cats" metaphor is used frequently to describe work with panels of expert clinicians, many of whom forcefully articulate strong and sometimes contradictory opinions. Nonetheless, their advice is invaluable and generally worth seeking. To identify candidate risk factors, expert panels must remain focused on the specific outcomes and time frames of the intended analysis. Frequent reminders to discriminate among specific risk factors relevant to different conditions or outcomes are important. To determine a parsimonious set of risk factors for a planned chart review tool, Daley and colleagues (1988) asked MMPS panelists to suggest small numbers of risk factors predictive of mortality across all four conditions and to propose several condition-specific variables.

Researchers should ask clinical experts to specify exactly how candidate risk factors relate to the outcome(s) of interest. For instance, some risk factors, such as body temperature, are abnormal at both high and low levels. How, therefore, should temperature be entered into a statistical model? Clinicians can clarify these relationships by drawing pictures or graphs. With increasing body temperature on the x-axis and increasing risk of death on the y-axis, the relationship between temperature and mortality risk roughly forms a U-shaped curve. This strategy worked well in developing the MMPS. Clinicians noted the critical importance of low body temperatures as well as fever in predicting the likelihood of imminent death among pneumonia patients. In the Surgical Risk Study, panels of clinicians reviewed all the candidate risk factors and succinctly summarized complex thinking by drawing sketches and graphs depicting the relationship between the continuous risk factors (e.g., laboratory variables, age) and the outcomes (e.g., U-shaped curves suggest elevated risk at high and low abnormal values).

Even as EHRs become increasingly available, translation of clinically important concepts into variables that can be measured reliably and validly using available data sources presents a significant challenge. The Surgical Risk Study began in 1991 after analyses of VA discharge abstract data and retrospective medical record reviews failed to provide findings with clinical credibility acceptable to surgeons practicing in VA facilities. Inconsistent coding across the 123 hospitals performing major surgery and the inability of discharge abstracts to distinguish preoperative from postoperative diagnoses (before present on admission flags) compromised the utility of the VA discharge abstract database (see Chapter 5). Retrospective chart review of six major surgical operations in a sample of VA medical centers demonstrated poor interrater reliability; furthermore, important risk factors were missing from the charts.

The VA and its surgeons decided that the only way to ensure reliability, validity, and clinical credibility was to have trained nurse reviewers concurrently gather data on major surgery patients during their hospital stays (i.e., prospective data collection). Project staff added the definitions and variable names to the charts of all surgical patients, and staff oriented all attending surgeons and surgical residents to the data collection procedures at each institution. Although the VA investigators probably could have gathered all preoperative and intraoperative risk factors from electronic records, they learned that reviewers needed to prospectively survey, identify, and correctly classify adverse postoperative events in a reliable way across institutions. The VA deemed that the direct costs of prospective data collection, analysis, and reporting—about $50 per case at the time—were worthwhile to produce comparative information across hospitals that chiefs of surgery would believe and therefore act on.

The eventual implementation of a system-wide computerized health information system across all VA hospitals facilitated electronic data collection and transmittal and fostered communication among sites nationwide. The NSQIP Surgical Package of the Veterans Health Information Systems and Technology Architecture (VistA, see Chapter 6) captured required data elements and enabled range checks, verification, and reporting of surgical volume, risk assessment, and operative information at each site. In 1998, NSQIP cost $38 for each major surgery assessed by the program, or $4 million across the VA system (Khuri et al. 1998, 499).

Researchers may not appreciate fully the challenges of gathering risk factor data until they try it. In studies using retrospective record reviews, even when reviewing EHRs, researchers should first examine five to ten records to determine their level of detail and completeness. Narrative information in EHRs is not necessarily more complete and accurate than in paper records, although the benefits of legibility are enormous. The common practice of copying and pasting portions of EHRs from one encounter record for a given patient to another encounter record may perpetuate documentation errors.

Although prospective data collection is generally more reliable and valid than retrospective record reviews, pretesting of data collection procedures and protocols is still necessary. In the Surgical Risk Study, pilot data collection of wound, blood, sputum, and urine culture bacteriology results revealed that the hospitals reported culture results very differently. This variability precluded collecting these data even prospectively; now these data can be downloaded electronically from VistA. Thus, institutional differences in the way standard tests are handled may compromise the use of results data in multicenter projects.

Lack of standardization among common diagnostic procedures may also complicate detection of risk factors. In developing the MMPS pneumonia models, clinical experts suggested that the number of lobes showing an infiltrate on a chest radiograph is a risk factor for imminent death. MMPS

investigators were unable to collect this information routinely from the chest x-ray reports for two reasons. First, among the targeted sick elderly population, many chest radiographs were portable anteroposterior films in which the number of lobes could not be assessed. Second, even among patients who had posteroanterior and lateral films, the radiologists' readings rarely noted the number of lobes involved with pneumonia. Although the patients' physicians may have known the number of lobes involved by physical examination or through their reviews of the radiographs, this information was not contained in their narrative reports.

The Surgical Risk Study restricted patient risk factors to those typically recorded in routine preoperative assessments and laboratory evaluations. As in developing MMPS, a six-hospital pilot study confirmed that collection of electrocardiogram (ECG) results was too time consuming. Laboratory values were collected by software programs that automatically scanned laboratory databases in each hospital and transmitted the information for analysis. In this large-scale observational study, the researchers avoided using certain cardiac, pulmonary, and vascular test results as candidate risk factors for several reasons: high rates of missing values (i.e., a small fraction of patients undergo these tests), inconsistent interpretation of results across facilities, and unacceptable data collection burden.

Finally, risk factors should reflect only patient attributes, not some component of processes of care, such as performance of a highly sophisticated diagnostic test available only in certain health care settings. This issue, however, highlights the complex trade-offs made in designing operational risk adjustment methods: Researchers may balance conceptual purity against practical results. For example, the VA chiefs of surgery strongly believed that surgical type and emergency status must be considered when assessing patients' risks for mortality or morbidity. They invested considerable effort in quantifying surgical complexity (Khuri et al. 1997, 317):

> To account for differences in the complexity of operations performed between medical centers, groups of subspecialists were asked to rank the complexity of each index operation in their subspecialty *above and beyond the risk factors that patients would typically bring to the operation*, on a scale of 1 to 5. The average score for each index operation, identified by CPT-4 code, was used as a measure of the complexity of that operation. [Emphasis added.]

In early models, across the NSQIP database, the "operation complexity score" was the eleventh most important predictor of mortality and the third most important predictor of morbidity; "emergency operation" was the fourth most important predictor of both outcomes (Khuri et al. 1998, 499). Other NSQIP risk factors, such as steroid use, blood transfusion, and ventilator dependence, also carried treatment connotations. ASA class, which is a subjective decision by the anesthesiologist, is another complexity example.

Using risk factors related to processes of care could particularly confound comparisons by hospital teaching status and size. Not surprisingly, VA teaching hospitals see patients with statistically significantly higher operation complexity scores than do VA nonteaching hospitals (Khuri et al. 2001, 373). Teaching-hospital patients generally had much higher rates of serious risk factors than did those at nonteaching hospitals. Nonetheless, after risk adjustment, 30-day postoperative mortality rates did not vary across teaching and nonteaching hospitals. Risk-adjusted 30-day morbidity rates, however, were significantly higher at teaching than nonteaching hospitals for patients undergoing general, orthopedic, urology, and vascular surgery operations.

Timing of Risk Factors

NSQIP began as a major initiative to improve surgical outcomes in VA hospitals. As noted earlier, to isolate that elusive quantity—quality of care—NSQIP predictive models of 30-day postoperative mortality and morbidity controlled only for preoperative findings. If intraoperative or postoperative factors had been incorporated, processes of care, and thus quality, could confound the results. Nonetheless, when evaluating postoperative length of stay (LOS) as an outcome, the NSQIP researchers considered a range of factors from the entire hospitalization. For major elective surgery, important risk factors for predicting postoperative LOS included older age; nonwhite race; ASA class 3; partially dependent functional status; intraoperative blood transfusion; operative time of three or more hours; postoperative urinary tract infection, ileus, or pneumonia; return to the operating room; and complication counts of two or more or three or more (Collins et al. 1999, 254). The researchers concluded, "Although preoperative factors were independently associated with a prolonged LOS, the factors generating the highest risks for a prolonged LOS were the intraoperative processes of care and postoperative adverse events" (Collins et al. 1999, 251).

A related topic is the time window for measurement of pertinent risk factors. For example, APACHE aims to predict in-hospital death. Because most physiological parameters are measured repeatedly in ICUs, numerous values for APACHE's risk factors are typically available, especially in electronic health information systems. Which values of these physiological parameters should be used for risk adjustment? The first? The worst over some period? What period?

The answer to this question depends on the context (see Chapter 4). Nonetheless, the decision has implications for data collection. Although APACHE's developers originally recommended using the most abnormal value from the first day of ICU treatment, studies of cases at their own institution showed that in 88 percent of the physiological measurements, the worst value was that observed on ICU admission (Knaus et al. 1985, 825). Depending on the time interval considered, worst values might reflect the

results of therapeutic or diagnostic mishaps. For example, blood pressure may become further deranged from its admission value because of incorrect therapeutic decisions. In addition, during manual reviews of EHRs or paper records, the first value is easier to identify than the worst value. Reviewers require less training to identify the first value, and the reliability of data abstraction is probably higher. EHRs might graph some laboratory results, however, which would help reviewers identify worst values. Nevertheless, extensive analyses by APACHE's developers showed no significant difference between APACHE risk models based on physiological parameters collected initially and those based on physiological parameters gathered within the first 24 hours. However, daily calculation of APACHE prognostic estimates during an ICU stay usefully tracks patients' clinical trajectories, differentiating patients who are improving from those who are worsening. Electronic availability of risk factor information facilitates the use of values from different periods.

The VA Surgical Risk Study initially assessed risk using 65 factors measured prior to surgery. Preoperative risk variables were captured as closely as possible to the time patients entered the operating room, but the investigators used some information from up to 14 days prior to surgery. Longer preoperative windows were especially common for low-risk, elective surgeries. For example, while the investigators wanted the most recent hematocrit and white blood cell count values, they needed to accommodate elective surgery; patients underwent ambulatory preoperative clearance sometimes up to two weeks prior to their operations.

The longer the period from which data are collected, the greater the possibility of confounding the quality of care with patient risk before treatment. If the period is too short, however, the amount of missing information may be unacceptably large. One study to predict in-hospital mortality at a tertiary teaching hospital used information from a computerized repository of all laboratory test results obtained on hospitalized patients (Davis et al. 1995). To avoid confounding quality with patients' risk factors, the researchers initially intended to include only laboratory values from the first 12 hours of the hospital stay; they had assumed that resident physicians in this intense academic medical center thoroughly worked up patients immediately upon admission. However, to their surprise, they needed to expand the time window to the first 48 hours of hospitalization. Otherwise, too many values were missing, even for routine laboratory tests. Continuing compression of hospital LOS may change the timing of testing and the availability of certain test results during the hospitalization.

Other Feasibility Considerations

Some variables are such important risk factors that one may reasonably devote considerable time and effort to collecting them. In contrast, marginally useful risk factors may not be worth expending energy to collect. In

developing the MMPS, Daley and colleagues (1988) conducted feasibility studies to determine the most reliable sources of risk information in charts. On the basis of these studies, they created a hierarchy specifying where data abstractors should look to obtain particular pieces of information. For example, because they sought the first vital sign values obtained in the hospital, they looked at emergency room information first and then reviewed the initial admission history and physical. Next they examined all progress notes recorded in the first 24 hours. In teaching hospitals, they abstracted information from the attending physician, resident, intern, and medical student, in that order.

The MMPS researchers also collected variables they considered important but difficult to abstract: functional status and do not resuscitate (DNR) status. Functional status information was available in only two-thirds of the charts and was rarely noted by the physicians. However, many of the initial nursing admission notes contained basic information about patients' capacity to ambulate and feed themselves. They therefore used nurses' notes as the source of functional status information. They collected DNR status only if it was present on admission. DNR status usually appeared on the emergency room admitting sheet and in the admitting orders, but it was designated differently across institutions (e.g., DNR, do not resuscitate, care and comfort only, no code 99, no code blue). Therefore, they specified synonyms that qualified as DNR orders. DNR status is located in different sections in different EHRs, so again reviewers had to know where to look for this variable.

As suggested by the Surgical Risk Study experience, prospective data collection may solve some problems relating to variable definition, standardization, reliability, and validity. Data collection instruments and guidelines are written in advance, and data collectors are taught and tested during a training period. Pilot tests of data collection procedures may determine the feasibility of collecting specific data elements.[2] Certain important risk factors represent sensitive information that is difficult to obtain reliably regardless of how data are gathered. For example, although sexual behaviors and use of illicit drugs are associated with selected outcomes, this type of sensitive information is often underreported in medical records. In prospective data gathering, respondents often refuse to answer, lie, or are so offended by being asked such personal information that they refuse to answer subsequent questions.

Building the Risk Adjustment Model

A combination of clinical judgment and empirical modeling produces better models than does either approach alone. Different risk adjustment methods have emphasized different development approaches, generally depending on whether large data sets are available for empirical modeling. Especially in the

late 1970s and early 1980s, few large data sets, even administrative files, were available. Methods dating from that era often relied primarily on the normative judgment of expert clinicians; empirical testing was limited. Nowadays, the widespread availability of large electronic files, even those containing extensive clinical information, offers opportunities for deriving both clinically credible and statistically rigorous risk adjustment methods. As noted earlier, this abundance of data poses a strong temptation for data dredging—a poor strategy for developing valid risk adjusters. The growing sophistication of statistical methods also requires developers to carefully choose their modeling strategy and validation techniques (Shahian et al. 2004, 1870):

> Risk model development requires considerable statistical expertise and judgment, a caveat that is sometimes forgotten in this era of ubiquitous, powerful, off-the-shelf statistical software. For example, the type of modeling strategy and validation techniques may differ depending on whether the purpose of the model is description of relationships (ie, comparison of providers or treatments) or prediction of future events.

Clinicians can help identify candidate risk factors and hypothesize about their relationships to the outcome(s) of interest. Without empirical input, however, clinicians are generally unable to quantify precisely the effects of risk factors, but statistical methods alone may yield clinically suspect results. In the Surgical Risk Study, one clinical variable—smoking in the two weeks prior to major surgery—demonstrated an unexpected statistical relationship to mortality. Patients who reported not smoking in the two weeks prior to surgery had much higher mortality rates than those who did report smoking, a relationship opposite of the one predicted. The researchers therefore compared other risk factors of nonsmokers and smokers and discovered that nonsmokers had much higher risk profiles (e.g., more diabetes mellitus, ischemic heart disease, chronic pulmonary disease) than those reporting smoking in the two weeks prior to surgery had. Contrary to the initial hypothesis, nonsmokers appeared very sick, perhaps physically unable to smoke in the two weeks before surgery. The researchers did not include this smoking variable in their final risk adjustment models; instead, they established smoking status by determining whether patients had smoked in the year prior to surgery and how many "pack-years" they had smoked.

Assembling an Analytic Data Set

Development of risk adjustment methods requires a data set, unless the method's specifications derive solely from clinical judgment. Often data must be "cleaned" before they can be analyzed (Cody 2008). Standard data cleaning involves range checks (e.g., looking for values that are not plausible, such as centigrade temperatures of 102°), identifying impossible occurrences (e.g., female patients admitted for prostate surgery, although characteristics of transgender individuals sometimes complicate seemingly straightforward gender

matches), searching for invalid data elements (e.g., codes that have no meaning in the coding nomenclature), and describing the frequencies of missing or poorly specified data elements (e.g., frequency of missing values for liver function tests in patients undergoing cholecystectomy). Multivariable checks should also be analyzed: Is the systolic blood pressure always higher than the diastolic blood pressure? Is the most extreme or worst value during the first 24 hours of admission always higher or lower than the first recorded value of the risk factor? Analysts acquire a valuable in-depth knowledge of their data sets through this process.

The extent of cleaning and editing a data file requires depends on the source of the data. In general, more cleaning is needed when the data have not been used for a similar purpose. Some analysts obtain administrative data files that have already gone through extensive checks. For example, the State Inpatient Databases, the core data set used in AHRQ's Healthcare Cost and Utilization Project (HCUP), undergoes a comprehensive set of cleaning edits. In its HCUP Quality Control Procedures, AHRQ (2008, 2) explicitly states its quality control philosophy:

- Make the data usable without extensive further editing.
- Confirm that data values are valid, internally consistent, and consistent with established norms, when feasible.
- Use some edit procedures to set questionable and inconsistent values to inconsistent . . . Use other edit procedures only to tabulate edit failures. Use the latter to evaluate whether systematic problems exist.
- Never "fix" or impute data. Set invalid or inconsistent values to special missing values. This preserves the analyst's ability to investigate data anomalies.
- Some data elements are more important than others because:
  - —they are coded more reliably because they relate to reimbursement; and
  - —without these data elements, a discharge record is not useful for most analytic purposes.

  Therefore, values of these data elements should be retained even in the presence of conflicting information. In order of importance, these data elements are:
  1. Discharge date (and within discharge date: year, month, and day)
  2. Admission date
  3. Principal diagnosis
- Tabulate instances of edit failures and use these to assess data quality for each data source.

If the data were entered using computer software that includes internal range checks, implausible values may have been rejected as the data were gathered. In developing the MMPS, Daley and collaborators (1988) created data-entry software that had internal data consistency checks and ranges of permissible values, thus preventing the entry of illogical or invalid values for many parameters. Some data files from chart review or prospective data

collection may need considerable cleaning. When data are acquired from external sources, it is often necessary to clarify variable definitions, determine coding conventions, examine the extent of missing data, and organize the data set for analysis. Clinical consultants may be especially helpful with this process. For example, clinicians can specify ranges for clinical variables that are physiologically implausible and thereby identify data errors.

Treatment of Missing Values

In most data sets, values of some variables are missing. This problem is found most frequently among data collected from retrospective medical record reviews, but administrative databases and even prospectively gathered information may be missing numerous values. Careful thought must be given to decisions on how to handle missing information. The risk adjustment literature addresses missing data concerns most commonly in the context of acute physiological parameters. Typically these variables are clinical guideposts for physicians caring for acutely ill patients, and most (e.g., vital signs, complete blood counts, serum chemistries) are measured routinely and through minimal technological intervention (e.g., venipuncture). One reason certain variables were eliminated in creating APACHE II was that they were measured infrequently (e.g., serum osmolarity, lactic acid level, skin testing for anergy), whereas others reflected a treatment decision (e.g., right atrial pressure measured through a central venous pressure line instead of through ultrasound). The remaining variables, which are generally available and easily measured, had good ability to predict in-hospital death (Knaus et al. 1985).

Given that most of these physiological parameters are measured routinely, how should one interpret missing data? *A priori*, this question has no correct response; it must be answered in both research and clinical contexts. For example, APACHE assumed that unmeasured parameters were likely to be normal. This assumption is reasonable, as APACHE pertains explicitly to ICU patients who are generally aggressively treated and monitored. The MMPS also substituted normal values for missing values after analyses suggested that this choice maximized predictive accuracy and reliability (Daley et al. 1988).

The approach toward missing values, however, may differ in other clinical contexts. It is therefore important to understand why values are missing. A recent Australian study aimed to create a parsimonious perioperative mortality score using a small set of variables: Age, albumin, and ASA score would serve as the main variables, and risk assessments were modified if patients required unplanned ICU admissions or developed inflammation (high white blood cell count) or renal failure (elevated creatinine) (Achuthan et al. 2011). To test the feasibility of creating these scores, the investigators audited records to see how often albumin and pre- and postoperative white blood cell counts and creatinine levels were measured. Only 47 percent of

patients had these basic tests performed. The total cost of testing all untested patients would be $18,927 (Australian dollars) per 1,000 patients.

Patients' preferences may limit testing and thus available data. An important example involves gravely ill patients, such as those with widely metastatic cancer, who explicitly refuse even routine monitoring by blood tests and request "comfort measures only" (e.g., pain medication easily administered through patches placed on the skin). In this context, absence of information about serum electrolytes or hematological indices reflects patient preferences, not physiological status. In still other clinical contexts, data are missing when patients die before tests can be performed. This situation may explain Blumberg's (1991) classic finding that AMI patients with missing laboratory values were more likely to die within 30 days of admission than were patients with complete information on routine tests.

Depending on the database used to assess acute physiological parameters, two further explanations for missing values require consideration. Practice patterns vary—even routine physiological information collection practices. In the United States, unnecessary and excessive preoperative testing is a long-recognized problem (Katz et al. 2011). Records in regions or facilities with heavy testing patterns contain results that are missing from documentation in settings with lower testing rates. Data collection protocols may be another reason data are missing. For example, the instructions in an old version of MedisGroups directed reviewers not to record values of clinical findings if they were in a "normal" range. Although this strategy aimed to facilitate the abstraction of literally hundreds of data elements, it produced problematic results. When MedisGroups KCFs were missing, analysts could not determine whether the test was not performed or whether the test result was normal. Furthermore, the normal ranges were broad (e.g., systolic blood pressures were recorded only if less than 90 mmHg), so many cases were missing values (Iezzoni et al. 1993). Later versions of MedisGroups gathered all values, not just those outside the defined normal ranges. Downloading these data directly from laboratory information systems increases the feasibility of gathering all laboratory values.

Some information is frequently missing but has important predictive value as a risk factor. For instance, in the MMPS data set, only two-thirds of the patients had a simple measure of functional status (ability to ambulate independently) recorded in their charts, usually in the nursing notes (Daley et al. 1988). The MMPS algorithm included this variable because it proved to be a significant predictor in some models. The researchers assumed that the absence of functional status information implied full ability to ambulate independently and substituted "fully independent" for the missing values.

In addition to the clinical and data collection implications of missing values, the way missing values are handled may dictate the number of cases available for analysis by some modeling techniques. For example, computerized

statistical routines often drop cases with any missing value among any of the variables in the model. Even if each variable in a model is missing values for only a few cases, the cumulative number of cases missing at least one value could become large if the model considers many variables. Thus, numerous cases could be eliminated during statistical modeling in data sets with frequently missing values. This elimination may bias the models toward characteristics of the subset of patients with complete data, who are unlikely to be a random sample. Thus, a model with few explanatory variables may be fitted to a larger data set than another model with many explanatory variables because the data supporting the second model would have eliminated cases that are missing values for any of its explanatory variables (i.e., the larger the number of explanatory variables, the larger the opportunity for missing values for some of these variables). To better understand the effects of missing values, analysts should use the same standard data set when comparing models built with different variables.

Given the potential confounding of patient risk with data quality, availability, and practice patterns, the way one treats missing data is important. In some contexts, substitution of normal values for missing values of clinical measurements is reasonable under the assumption that abnormalities would be recorded in the medical record. Daley and colleagues (1988, 3620) took this approach in devising the MMPS, justifying this decision by observing that "studies . . . on other data sets indicate that this rule maximizes predictive accuracy and replicability." This approach is an example of *single imputation*—that is, substituting a single replacement value for a missing value. For variables for which no natural normal value (e.g., education) exists, the substituted value might be the mean of the observed values for all people or subsets of people with similar characteristics to those whose values are being imputed. (This mean might be estimated by using a regression model.) When using single imputation, one performing a standard analysis[3] of the data set must explicitly take into account the increased uncertainty introduced by using imputed values; otherwise, standard errors associated with estimates from the data set will be too small.

Multiple imputation addresses problems presented by single imputation (Little and Rubin 1987; Schafer 1999; Patrician 2002; Donders et al. 2006; Wang and Wu 2011). This approach replaces each missing value with a set of $m > 1$ plausible values drawn from a probability distribution of the missing value. The result is m complete data sets (usually three to five imputed data sets are sufficient), which are then analyzed in the standard way. The variation among the m data sets reflects the uncertainty with which missing values can be predicted from observed ones. Rubin (1996) shows how to combine results from the m data sets to obtain overall estimates and standard errors. Several sources offer computer software that imputes values for missing data.

In the Surgical Risk Study, all patient preoperative risk factors were more than 99 percent complete, except the preoperative laboratory values. Most of the preoperative laboratory variables (e.g., complete blood count, serum sodium, BUN) were more than 95 percent complete; a handful of preoperative laboratory values (serum albumin, serum bilirubin, serum glutamic-oxaloacetic transaminase) were missing in a substantial minority of cases (e.g., serum albumin was missing for 39 percent of cases). Analyses demonstrated that patients missing albumin and liver function test results were less sick (i.e., had a lower incidence of other risk factors) and had lower mortality and morbidity and shorter LOS than did patients with values for those laboratory tests. Extensive analyses compared incorporating imputed missing values using a regression procedure with substituting a normal value for each missing laboratory variable. These comparisons found similar results for the two approaches.

In certain contexts, especially aggressive public provider report card initiatives, the way missing values are handled could have practical consequences. If missing values are assumed to be normal, clinicians will learn that, to have their patients classified as higher risk, they must record the problems they find. This incentive could affect testing practices (e.g., increased laboratory testing even when clinical suspicions of an abnormal result are low).

Structure of Continuous Independent Variables

It is important to explore the effect of the form of each risk factor—whether it is entered into the model as a categorical or continuous independent variable. The most common continuous risk factor, age (see Chapter 3), is a case in point. The relationship of age to outcomes (e.g., in-hospital mortality, LOS, charges) is unlikely to follow a simple straight line, especially over a wide age range. Age can be treated as a continuously valued variable or categorized into two or more groups. For example, PRISM weights physiological values differently for two age levels, infants and children, thus incorporating an interaction between age and physiological status (Pollack, Ruttimann, and Getson 1988). APACHE III assigned points for patients in different age categories (Knaus et al. 1991); APACHE IV modeled age grouped into eight categories (the data set excluded persons younger than age 16) and several terms from spline functions (Zimmerman et al. 2006). Different risk adjustment measures model the effects of age with regard to outcomes in different ways.

Other clinical variables, such as blood pressure and temperature, have complex relationships to various outcomes, including death. For instance, the high and low extremes of blood pressure are associated with increased likelihood of imminent death as compared to the middle or normal ranges. However, although high blood pressure is a risk factor for long-term mortality, it may not be highly predictive of 30-day mortality; low blood pressure is associated with short-term mortality. U-shaped relationships with mortality may not

be symmetrical for high and low values of some variables. For instance, the likelihood of death may be much higher for incremental decreases in temperature and blood pressure than for similar incremental increases above normal values. In pneumonia cases, for example, Daley and colleagues (1988) modified the APACHE II scoring system for the MMPS to account for the much increased likelihood of death among the Medicare elderly presenting with very low body temperatures.

Most clinical variables have a nonlinear relationship to the outcome. In general, clinical judgment combined with statistical analyses is used to determine ranges for continuous variables and relationships to the outcome within these ranges. Computer-based "smoothing techniques" can help delineate these relationships. "Cut points" (i.e., points at which a large change in outcome occurs as a result of small changes in independent variables) in the smoothed function can be used to define categories of independent variables, which are then represented in the final model by dummy variables. Cubic splines are also used to produce smoothed functions (Harrell, Lee, and Pollock 1988). For example, the developers of APACHE IV did the following (Zimmerman et al. 2006, 1299):

> For each patient we estimated the probability of hospital mortality using a multivariate logistic regression procedure. . . . The age, APS [acute physiology score], and prior length of stay variables were allowed to have nonlinear relationships with outcomes using restricted cubic regression splines. Splines allow estimation of a nonlinear relationship between a variable and an outcome and replace less accurate techniques that assume the relationship is linear. Cut points (knots) are chosen and a separate coefficient is included for each interval between knots. A restricted cubic spline transformation was used to expand age and the APS to five nonlinear terms and previous length of stay to four nonlinear terms.

In the Surgical Risk Study, panels of clinicians were asked to sketch the relationship between the outcomes of interest and the continuous variables of age and preoperative laboratory tests. Age and serum albumin, empirically and by clinical judgment, demonstrated a linear relationship to the outcomes. Albumin produced a flat relationship when greater than 3.5, but below 3.5 the relationship was an exponential function. Age was linear in all the models. Selected laboratory variables (e.g., serum sodium, serum potassium, white blood cell count) demonstrated a U-shaped relationship to the outcomes and were trichotomized for the final analyses.

A continuous variable that appears in many disease-specific mortality prediction models is the serum BUN. In healthy patients, BUN ranges from 2 to 20 mg/dL, but in patients with multiple cardiac, renal, and metabolic abnormalities, BUN may rise to 500 mg/dL. In the MMPS, BUN was transformed and standardized to values greater than 40 mg/dL, and a continuous function of BUN values above 40 mg/dL was constructed. In the VA Surgical

Risk Study, a similar relationship was found, and BUN values greater than 40 mg/dL were used in all of the risk models.

Need for Data Reduction

Despite the many data limitations described in chapters 5 through 7, much useful information is nonetheless available for deriving risk adjustment models. The quantity of information can be overwhelming, and researchers may wish to reduce the data set to a reasonable number of potential risk factors for modeling, as suggested in Exhibit 8.2 (the NSQIP parsimonious models).

An example of the potential perils of incorporating too many predictors was demonstrated by a study that modeled in-hospital mortality for 16,855 patients aged 65 or older discharged from 24 MedisGroups member hospitals (Iezzoni et al. 1992b). My colleagues and I examined several risk models (Exhibit 8.3). The first model was the old five-level admission MedisGroups score (a categorical variable with a value of 0 through 4). The second and third models were empirically derived using KCFs specified for each condition. From the more than 500 potential KCFs contained in that early version of MedisGroups, for each condition we identified KCFs that were

Condition	Sample Used for Model Performance	Model Performance (R^2)		
		Admission MedisGroups Score	Top Ten KCFs	All KCFs
Stroke	Fitting	0.122	0.299	0.330
	Cross-validating	0.096	0.233	0.208
Lung cancer	Fitting	0.049	0.280	0.344
	Cross-validating	0.058	0.242	0.054
Pneumonia	Fitting	0.137	0.247	0.278
	Cross-validating	0.154	0.243	0.240
AMI	Fitting	0.204	0.243	0.276
	Cross-validating	0.195	0.180	0.157
CHF	Fitting	0.069	0.208	0.234
	Cross-validating	0.075	0.196	0.196
All cases	Fitting	0.134	0.260	0.293
	Cross-validating	0.133	0.225	0.195

Source: Adapted from Iezzoni et al. (1992b).

EXHIBIT 8.3 Fitting and Cross-Validating Model Performance in Predicting In-Hospital Mortality (R^2)

present in at least 1 percent of the cases or that were viewed clinically as important predictors of death (e.g., ventricular fibrillation). We identified between 40 and 65 KCFs for each of the five conditions. As described earlier in the chapter, the second model used forward stepwise regression to select the ten most statistically important KCFs following age and sex for each condition. The third model used all 40 to 65 KCFs plus age and sex for each condition.

We used the R^2 statistic from an ordinary least squares regression to compare the utility of the different models in predicting death (see Chapter 10). We first computed R^2 values on half of the data used to derive the models and then validated on the remaining half. The two empirically derived models had the best model fits (i.e., highest R^2 values). Although the model using all KCFs fit the development data somewhat better than the model using only ten KCFs, its cross-validated performance was never superior to that of the ten-KCF model. For lung cancer (total n = 1,244; 23.9 percent in-hospital deaths), the drop for the all-KCF model from the fitting R^2 value (0.344) to the cross-validated value (0.054) was particularly striking. Thus, although the models with 40 or more explanatory variables each achieved the highest R^2 values, these models appeared to have been overspecified; in the cross-validation analyses, they performed no better, and in one instance performed far worse, than the models based on the ten most important clinical variables (Iezzoni et al. 1992b).

A variety of approaches are available to trim the number of predictor variables prior to modeling. For example, as in the study just described, inspection of the frequency distributions of the variables may reveal items that appear too infrequently to be retained in the analysis. Removal of variables that are unreliable or of suspicious quality is prudent, not only because it eliminates concerns that they will introduce statistical error but also because their inclusion might produce models that are conceptually problematic in a clinical sense. A common method of removing variables is to examine bivariate associations between individual predictors and the outcomes. Model developers then employ only factors that are statistically significant at a prespecified level (e.g., $p < 0.10$).

The use of rare risk factors as predictors raises two distinct concerns. One is that too few cases are present in the model development data set to accurately assess the association between the rare risk factor and the outcome. For example, in a data set of 3,000 cases, a risk factor present in fewer than 0.5 percent of the cases occurs fewer than 15 times. This problem is minimized by a sufficiently large data set for model fitting. The other concern is practical. If data are gathered through manual chart review—even manual review of EHRs—do developers want to expend costly effort finding rare risk factors? The answer may be yes in three instances: First, the rare risk factor is strongly associated with the outcome; second, the rare risk factor

independently predicts the outcome even after other variables are included in the model; and third, clinicians feel that eliminating it substantially reduces the model's clinical credibility or penalizes certain providers who see a disproportionate number of rare cases.

As suggested earlier, statistically significant bivariate associations between risk factors and outcomes that appear opposite to the clinically hypothesized relationships must be carefully reviewed. They may reveal problems with the original data sources, the database itself, or the coding of the independent variables. A variety of computer-intensive approaches are increasingly used to identify important variables to include in models. Normand and colleagues (1996) used such an approach to identify candidate variables for a logistic regression model predicting 30-day postadmission mortality of AMI patients. From an initial sample of 14,581 patients, three 25 percent subsamples of patients were selected. For each of the three subsamples, 20 "random starting models" were identified. In each of the 60 models (three subsamples multiplied by 20 starting models), four clinical variables closely related to left ventricular ejection fraction were forced into the model. Using a stepwise procedure, other statistically significant independent variables were entered into the model. For each of the three subsamples, the best-fitting model among the 20 random starting models (measured by the likelihood estimate) was selected. The set of variables from these three "best-fit" models then became candidate variables for further regression models.

There are formal analytical techniques for data reduction, such as cluster analysis and factor analysis. However, these techniques have rarely been used in building risk adjustment models in part because clinical interpretation of the composite measures is difficult. As discussed earlier, if clinicians do not feel comfortable with the resulting models, much of the models' value is lost.

Multivariable Modeling Techniques

Detailed descriptions of multivariable modeling techniques, their underlying assumptions, strengths and weaknesses, and appropriate diagnostic measures are beyond the scope of this book. Readers should consult standard texts of statistical methods for detailed discussion of approaches to building multivariable models. In large part, the availability of clinical knowledge about the relationship of independent variables to outcomes is the main factor that distinguishes risk adjustment modeling from general multivariable modeling.

Chapter 10 discusses two main types of multivariable models used in risk adjustment: multiple regression models (used when the outcome is continuous) and logistic regression models (used when the outcome is dichotomous). In addition, Chapter 10 briefly discusses proportional hazard models, which are used when the outcome is time to an event, such as death. Researchers have evaluated a variety of alternative modeling approaches, particularly for

dichotomous outcomes. Logistic regression models have been shown to perform better than these alternatives.

Researchers often use stepwise procedures to build all three models. Briefly, forward stepwise regression procedures build the model by adding one variable at a time. At each stage, the variable added is the one that contributes the most to model fit at that step. The backward elimination approach begins with all variables in the model and then, one by one, eliminates the variables that contribute the least to model fit. Each technique includes options to drop variables previously added (forward elimination) or to add variables previously dropped (backward selection). In addition, some "stopping rule" is necessary. Usually modeling is considered complete when the variables added to the model are statistically significant at $p < 0.05$ or 0.10 (or those dropped are insignificant at one of these levels).

Harrell and colleagues (1984) performed simulations demonstrating that, if logistic regression models are built using stepwise procedures on data sets of fewer than 1,000 observations, the c-statistic (see Chapter 10) may significantly deteriorate upon validation. With a dichotomous outcome such as mortality, the number of cases in the smaller of the two groups (i.e., those with or without the outcome) is usually the limiting factor. The work of Harrell and collaborators (1984) suggests that first clustering the variables and then developing indexes from each cluster (using all the variables, a subset of variables, or both) performs better than do traditional stepwise procedures. Most risk adjustment models have been developed on data sets of more than 2,000 cases. When built on large data sets, models developed using stepwise procedures seem to validate as well as those developed using other approaches.

A major challenge in developing a model is to identify important interactions, or nonadditive effects, among predictor variables. Even when a moderate number of predictors are under consideration, possible interactions are generally too numerous for unguided statistical exploration to detect the important ones. For example, ten predictors generate 45 possible paired interactions and 120 three-way interactions. Knaus and colleagues (1991) used logistic regression results plus clinical judgment to study interactions among physiological variables in APACHE III, evaluating both individual and combined weighting of variables. An important example involved the variables reflecting clinical acid-base disturbances (i.e., serum pH, pCO_2, and bicarbonate). Using their database of more than 17,000 ICU admissions, they found empirical relationships incompatible with established physiological principles (Knaus et al. 1991, 1621):

> The computer-derived weights for serum pCO_2 above 50 mm Hg were consistently estimated as having little or no significant relationship to risk of death. We hypothesized that this was because the appropriate weighting for pCO_2 is also dependent on the

associated serum pH (i.e., whether there is a primary or secondary respiratory disorder). Therefore, we developed a combined variable, which included serum pH and pCO_2 to establish weights for common acid-base disorders.

They then derived the weight for this combined variable as they had for individual variables. They also found important statistical interactions between urine output and serum creatinine and among respiratory rate, $PaCO_2$, and ventilator use. They created combined variables from each of these sets and then compared the clinical validity of the weights assigned to the combined variables to the clinical validity of the weights assigned to the individual variables.

As suggested earlier, a major concern in model development is model "overfitting"—including variables that may be useful predictors in the development database but do not have the same relationship to the outcome in other databases. The lung cancer example in Exhibit 8.3 exemplifies this problem (Iezzoni et al. 1992b); stepwise regression models using all MedisGroups KCFs did not validate as well as simpler models. Chapter 10 discusses these issues further. Briefly, two main tactics are used to guard against model overfitting. First, developers base their decisions about which variables to consider on clinical and statistical criteria, and second, they limit the number of candidate variables. For example, for predicting a continuous outcome variable, the number of predictor variables should never be more than one-tenth of the number of cases. The preferred ratio is 30 cases per predictor variable. For modeling dichotomous outcomes, the standard rule is to include no more than one predictor variable per 10 cases experiencing the outcome of interest (Shroyer, Grover, and Edwards 1998). Using no more than one predictor for every 20 positive cases is safer.

Researchers must decide not only the number of independent variables to include but also whether to transform them before adding them to the model. Again, a combination of clinical judgment and statistical criteria underlies this decision. For example, in developing the MMPS, Daley and colleagues (1988) used the first 600 cases to select variables and choose the form in which each variable was expressed (i.e., APACHE II points, logarithm of the variable, or some other transformation), employing goodness of fit measured by chi-square analysis. Using an additional 300 cases, they next tested a few models, examining the overall goodness of fit but not the coefficients of individual variables. Only after the final functional form of the model was produced did they determine coefficients from the entire sample.

After a model is fit, researchers often must turn its output into a scale or score. In some cases, the results of the logistic regression model (e.g., parameter estimates, intercept terms) are used as the risk score. Examples include the methods developed by Elixhauser and colleagues (1998) and HCCs (Pope et al. 2004) mentioned earlier.

Conclusions

Over the last few decades, research on the development of risk adjusters supports the use of empirical techniques whenever appropriate data are available. No matter how sophisticated, however, empirical methods are generally not enough to produce a clinically credible risk adjustment method; clinical judgment is also required. Even state-of-the-art statistical approaches and large databases may yield clinically implausible findings. A model that is not clinically sensible (i.e., that contradicts established physiological principles) is not valid regardless of the statistical rigor used in its derivation. Studies that use clinically credible risk adjustment strategies are far more likely to yield findings that will be trusted, believed, and, most important, acted on. Because assessment of health care outcomes ultimately aims to affect clinical practice, acceptance of the risk adjustment method by health care professionals is essential.

Validity is thus a crucial attribute that encompasses both clinical and statistical considerations. Chapter 9 focuses on validity measurement primarily from a clinical perspective, and Chapter 10 addresses statistical measures of predictive validity.

Notes

1. Version 3.2 of the AHRQ QIs uses as its reference population the 2003–2005 State Inpatient Databases (SID) from AHRQ's Healthcare Cost and Utilization Project (HCUP, see Chapter 5). SID includes hospital discharges from all payers, including Medicare, Medicaid, private insurance, and other sources, and the reference population includes discharge data from 38 participating states (AHRQ 2010a).
2. The pilot test of the Surgical Risk Study's prospective data collection procedures identified various logistical problems. At the time, most laboratory data were gathered through automated software that downloaded required information directly from the hospitals' laboratory systems. ECG results, however, were not available from the automated system. In a feasibility study, data collectors spent unacceptably long periods retrieving ECG results from patient charts and the ECG laboratories to the detriment of collecting other important data elements. In addition, despite clear definitions of what constituted a nosocomial infection (i.e., postoperative wound infection), ascertainment bias became evident across the sites because of different practices for obtaining wound cultures. Adoption of Centers for Disease Control and Prevention definitions for superficial and deep wound

infections and pneumonia and urinary tract infections, which are also used as the infection control definitions throughout the VA system, permitted standard case finding (i.e., finding cases experiencing the targeted condition).

3. Statistical models are built on a variety of assumptions, some of which are considered standard (e.g., assumptions about distribution of values of independent and dependent variables).

CHAPTER

9

VALIDITY AND RELIABILITY OF RISK ADJUSTMENT METHODS

Lisa I. Iezzoni

The meaning of the phrase "valid and reliable" seems self-evident, as does the need for risk adjustment methods to be valid and reliable. However, demonstrating validity and reliability in the context of risk adjustment is complex and sometimes elusive. Concerns range from the validity and reliability of the methods themselves to the attributes of the data used for risk adjustment to the ultimate purpose of the risk-adjusted findings. The important question is not "Is this risk adjustment method valid and reliable?" but "How believable and trustworthy are the findings when risk adjustment method X is used to answer question Y?" For example, when risk adjustment aims to facilitate comparisons of mortality rates across hospitals, under the assumption that hospitals with excessively high rates might have quality problems, determining the validity of the risk adjustment ultimately requires a valid measure of hospital quality. The catch is obvious: Truly valid measures of hospital quality are rarely available.

Validity and reliability are multidimensional concepts without clear boundaries. According to Donabedian (1980, 101):

> The concept of validity is itself made up of many parts; and there is no precise way of saying what belongs to it, or what belongs more appropriately under another heading. . . . I would say that the question of validity covers two large domains. The first has to do with the accuracy of the data and the precision of the measures that are constructed with these data. The second has to do with the justifiability of the inferences that are drawn from the data and the measurements.

Donabedian's observations highlight the difficulty of distinguishing attributes representing validity from those connoting reliability or some other concept. Data accuracy and measurement precision are good examples. For instance, as described in Chapter 6, the medical literature documents physicians disagreeing among themselves about the presence of physical examination findings and interpretation of diagnostic tests, even technologically sophisticated studies. Does this unreliability make physician examination findings and diagnostic test results invalid as risk factors? If the differences involve random errors, only accuracy is affected. If systematic differences of

policy relevance occur among the groups or populations characterized by such risk factors, validity is threatened.

An additional consideration in assessing the validity and reliability of risk adjustment methods is the advancement of clinical knowledge and changing therapeutic practices. Relationships between individual risk factors and clinical outcomes might change over time as new clinical insights and treatments develop. Depending on the clinical context, risk adjustment methods need to be updated periodically to reflect new clinical knowledge, new data (e.g., the results of new diagnostic tests), and new diagnosis or procedure codes (e.g., the switch to ICD-10-CM; see Chapter 5) and to address problems identified by previous users. As risk adjusters change, or simply as time passes, their validity or reliability may change.

This chapter explores how the concepts of validity and reliability apply to risk adjustment methods, drawing key examples from seminal risk adjustment studies. Although the basic concepts of validity and reliability do not change over time, methods of assessing the validity and reliability of risk adjustment methods will likely evolve in coming years, especially as clinical data from electronic health records (EHRs) and information obtained directly from patients become increasingly available (see chapters 6 and 7).

Dimensions of Validity

Evaluating the validity of a risk adjuster involves answering the following question: How well does the adjustment method account for the true risk of a specified outcome within a particular time frame for a particular patient population for a specific purpose? As noted throughout this book, embedded within risk adjustment methodologies are answers to four questions (see Chapter 2). Risk adjusters developed for specific outcomes, time frames, populations, and purposes are generally most valid when applied to projects with similar parameters. As an extreme example, a risk adjuster designed to predict newborn mortality in neonatal intensive care units would have poor validity for predicting annual costs for elderly Medicare beneficiaries in managed care.

Evaluating the validity of scales or classification systems falls within the purview of psychometrics or clinimetrics (Carmines and Zeller 1979; Feinstein 1987; Stewart, Hays, and Ware 1992; Streiner and Norman 1995; de Vet, Terwee, and Bouter 2003).[1] Methodologists distinguish numerous dimensions of validity, as shown in Exhibit 9.1. These dimensions overlap. For assessing risk adjustment methods, the most important dimensions are face validity, content validity, criterion or construct validity, predictive validity, and attributional validity. This discussion does not re-create detailed technical descriptions available in psychometric or clinimetric texts but highlights issues most crucial in risk adjustment.

EXHIBIT 9.1 Dimensions of Validity

Validity Dimension	Definition	Example
Face validity	A measure contains the types of variables that will allow it to do what it aims to do	A method for adjusting for in-hospital mortality for AMI includes clinical variables that on face value are the types of variables clinicians consider important risk factors
Content validity	A measure contains all relevant concepts	A method for adjusting for in-hospital mortality from AMI includes all clinical variables that are important risk factors
Construct validity	A measure correlates with actual indicators of risk in the expected way	A method for adjusting for in-hospital mortality from AMI correlates with actual measures of cardiac function
Convergent validity	A measure has a positive correlation with other indicators of actual risk	When a method for adjusting for in-hospital mortality from AMI shows increasing risk, actual measures of cardiac functioning also show increasing risk
Discriminant validity	A measure has a stronger correlation with indicators specific to its purpose than to other indicators	A method for adjusting for in-hospital mortality from AMI correlates more strongly with actual measures of cardiac function than with measures of ambulation
Criterion validity	A measure correlates with the gold-standard measure	A method for adjusting for in-hospital mortality from AMI correlates with a clinical scale derived from intensive, continuous cardiac monitoring
Predictive validity	A measure explains variations in outcomes	A method for adjusting for in-hospital mortality from AMI predicts accurately which patients have died
Attributional validity	Findings using the measure permit one to make statements about the causes of what is observed	In-hospital mortality rates, adjusted using the measure, permit one to attribute differences to effectiveness or quality of care

Face Validity

Face validity indicates whether a method appears "on its face" to measure what it claims to measure. In other words, would the method's users accept it as "valid" in the everyday sense of the word? Face validity, while not a rigorous technical concept, is critically important. Poor face validity impedes overall acceptance, especially by practicing clinicians with little knowledge of more methodological issues in risk adjustment. Clinicians generally judge the acceptability of a risk adjuster against their own standards: How well do the ratings match their intuitive sense of how sick or impaired a patient is?

Clinicians are skeptical of risk adjustment methods that diverge from their clinical expectations about patients' prognoses. As a narrow example, physicians expect patients with chronic renal failure requiring dialysis to face higher risks of complications or death when hospitalized (e.g., for major surgery) than they do patients who do not have end-stage renal disease or other major health problems. Risk adjustment methods that rely solely on serum BUN to capture renal failure effects may underestimate the likelihood of complications or death for patients on hemodialysis; dialysis lowers patients' BUN levels, which otherwise would be grossly elevated (i.e., a low BUN level does not indicate good renal functioning in dialysis patients). For risk adjustment in populations containing renal dialysis patients, methods must consider clinical risk factors beyond BUN to have face validity.

Examining the face validity of a risk adjustment method requires evaluating its inner workings. The logic of the method (including risk factors and weights) should be fully available for clinicians and methodologists to examine. Over the last two decades, the internal workings of many commercial risk adjusters have been kept confidential as trade secrets by their vendors. Nevertheless, to assess face validity, clinicians need to consider whether the risk adjusters include all important risk factors and whether the direction and weight of each seems appropriate for the outcome being predicted. As suggested in Chapter 8, risk factors that behave differently than clinicians expect raise questions about face validity and may cause clinicians to reject the risk adjustment method.

A key step in assessing face validity is perhaps the most difficult: translating the analytic structure of a statistical risk adjustment model into information accessible to clinicians so that they can test it against their clinical experience. This translation is especially difficult when empirical models include highly correlated factors, numerous interaction terms, and hierarchical structures. One complexity is determining the optimal representation of various independent or predictor variables in predictive models and the weight each variable should have. Graphing the predicted effects of various risk factors on the outcome may help but may not fully resolve questions about the optimal way to represent variables in predictive modeling.

Content Validity

Content validity refers to the extent to which the risk factors included in the method represent the universe of risk factors that should be incorporated. Models of risk can always include more factors (see Chapter 3). Examination of content validity requires determining whether important risk factors are missing. The information used to judge this dimension is typically drawn from the clinical literature and expert clinicians, as described in Chapter 8. Again, relevant risk factors depend on the outcome, time window, and population of interest.

Assessing content validity is complicated when the independent variables are highly correlated, causing some clinically important variables to be excluded. For example, in the national VA Surgical Risk Study introduced in Chapter 8, general surgeons examining the final risk models for 30-day mortality and morbidity often commented on the absence of key laboratory values, such as bilirubin and liver function tests, for assessing the presence of acute and chronic hepatobiliary disease. Initially, the surgeons questioned the validity of these models. After discussion, however, they appreciated that serum albumin level (an included variable) was an excellent predictor of surgical outcome, representing risk from both hepatobiliary disease and poor nutritional status. Although important variables may not appear in a final prediction model, other closely related variables may capture essentially the same information. Efforts to build parsimonious risk adjustment methods in the National Surgical Quality Improvement Program (NSQIP), the follow-up effort to the VA study (see Chapter 8), may raise questions about content validity, especially when models include only two to five variables (Dimick et al. 2010); however, the trade-off between the cost and feasibility of collecting extensive clinical data and the methods' capabilities (e.g., in terms of content validity, predictive validity) is clear.

As noted, an individual clinical parameter can capture a range of clinical concerns. For instance, serum BUN, which appears in many acute care risk adjustment models, reflects multiple clinical issues: renal insufficiency, intravascular volume depletion, diminished cardiac output, significant gastrointestinal bleeding, and an increased catabolic state. In contrast, several risk factors may be needed to fully represent some clinical risk concepts. For example, a highly valid representation of risk of death from sepsis or septic shock includes very low or very high body temperature, low blood pressure, very high or very low peripheral white blood cell count, and pathogen-positive blood cultures. No one or two of these variables alone have full clinical validity for predicting the risk of death from sepsis.

Nonetheless, content validity is ultimately limited by the information in the data source. While administrative data in particular contain limited clinical insight (see Chapter 5), key risk factors may be unavailable from any

practical source. This situation may change with increasing availability of EHRs, but regardless of the medium (electronic or paper) some clinical dimensions may not be fully represented in medical records (e.g., patients' social and environmental determinants of risk). The absence of routinely collected data on psychosocial functioning and quality of life limits the utility of risk adjustment methods for certain purposes.

Criterion and Construct Validity

Criterion validity is the extent to which a given measure correlates with a gold standard or criterion. While no gold standard exists for an abstract, multidimensional concept like risk, some validation procedures may be useful. *Construct* or *correlation validity* is verified by finding a strong, positive association with another credible measure, such as assessments of risk by expert clinicians. Sinuff and collaborators (2006) conducted a systematic review comparing the accuracy of predictions of hospital mortality for critically ill adults by ICU physicians in the first 24 hours versus risk-scoring systems.[2] They identified 12 observational studies that met their inclusion criteria. "Combined results of all 12 studies indicated that physicians predict mortality more accurately than do scoring systems: ratio of diagnostic odds ratios (95% confidence interval) 1.92 (1.19, 3.08) (p = .007)." Using data from seven studies, they found the c-statistic was "0.85+/−0.03 for physician predictions compared with 0.63+/−0.06 for scoring system predictions (p = .002)." They concluded that "ICU physicians discriminate between survivors and nonsurvivors more accurately than do scoring systems in the first 24 [hours] of ICU admission" (Sinuff et al. 2006). Physicians decide whom to admit to ICUs when beds are in short supply (Sinuff et al. 2004), so it is reassuring that ICU physicians predict mortality among critically ill adults relatively well. Nonetheless, the systematic review findings suggest that the criterion validity of the evaluated scoring systems could be improved.

Another method of assessing criterion validity is to compare a new risk adjustment approach to an established, well-accepted method. NSQIP investigators compared their model to APACHE. Noting the similarity of the clinical information used by each algorithm, they speculated that NSQIP data could produce valid predictors of mortality (Turner et al. 2011). Another way to investigate construct validity is to compare scores among risk adjustment methods (see the example from Shahian et al. [2010] in Chapter 2), but this comparison must be done using a single data source. Some data are more difficult to model than others. Thus, ideally, comparisons are based on the same patient population, outcome measure, and statistical approaches to assessing predictive validity (see the following discussion). If use of the data is impossible, careful attention to how the data sets differ with respect to distributions of the dependent and independent variables, as well as reviews

of performance characteristics of other models in similar settings, may help in interpreting results.

Predictive Validity

Predictive validity refers to how well a risk adjuster predicts an outcome. Chapter 10 discusses predictive validity in detail by examining statistical measures of model performance. This chapter addresses conceptual issues regarding assessment of predictive validity.

When considering predictive validity for an empirically derived risk adjuster, researchers need to distinguish between the model's performance with the data set used to develop or "fit" it and its performance with "validation" data. Not surprisingly, an empirically derived risk adjuster typically performs better with development data than with other data. However, the true test of a model's predictive validity is how well it works with new data.

Researchers typically create a specific, well-defined database (analytic file) for their studies. Administrative databases may contain hundreds of thousands or even millions of observations; data sets from chart review or clinical studies generally contain much fewer observations, typically only several hundred cases, although data downloaded from EHRs may contain many thousands of cases. Often researchers split their cases into two or more subsets, using one to develop their predictive model and another (or others) to test (validate) it. For example, Wong and colleagues (2011) used data from an EHR of almost 160,000 hospitalizations to develop models that could predict daily risk of death for hospitalized patients. They developed their model using time-dependent Cox regression methods and two-thirds of their data set (randomly selected), validating the models on the remaining one-third of the data. A similar project using EHR data from nearly 260,000 hospitalizations to develop risk adjustment methods for predicting mortality also performed split-sample validation (Escobar et al. 2008).

The more data used to develop the risk adjustment model, the greater the ability to identify important risk factors (predictor variables) and accurately quantify risks. Nonetheless, because all the data are not used in initial model development, analysts should report performance measures from both the split-sample models and the final models derived from the entire database. Although the validation sample results do not directly pertain to the model developed using the full sample, the full-sample model should perform even better than the validation measures suggest. When developing a model whose specific coefficients will be used repeatedly, analysts can better understand how well a full-sample model captures real and stable relationships by examining whether models developed using subsets of the data have the same coefficients and predictive validity. If subsample models display considerable variability, analysts may need to use additional data to establish fully

credible models. In databases of hundreds of thousands of cases, 100,000 cases should be sufficient for the validation sample. Analysts can also use bootstrapping to examine the stability of coefficients and measures of predictive validity for different data sets (Efron 1979; see Chapter 10).

As this discussion suggests, development of a risk adjustment method requires multiple steps, each guided by extensive examination of the development data. Critical steps include coding risk factors, selecting risk factors to include in the model, and assigning them weights. Each step presents an opportunity to "overfit" the data (i.e., to make a choice that is optimal for the particular data set) and thereby build a model that fits those data better than it can be expected to fit any future database. Overfitting produces inflated performance measures. The cleanest way to avoid being overly optimistic about how well a model will perform in the future is to look at the development data during the entire model-building process and use the validation data only at the end.

With limited data, analysts often circumvent the ideal situation of having entirely different, large data sets for model development and model validation; analysts who have only small databases, for example, may use the entire data set to identify the model's predictor variables and validate only the coefficients produced by fitting the model. However, methods that reduce the number of variables, such as stepwise selection, may yield unstable results. Repeating the entire modeling process with data subsets can expose the extent of that instability in a particular situation. To interpret cross-validated performance measures, one must know whether the entire model-building procedure or only selected components of the process were validated.

When possible, external validation with entirely independent data is best. However, if the mean of the outcome variable is substantially different in the independent validation data, researchers may want to consider at least some recalibration of the development model. Chapter 10 discusses these issues more fully.

Attributional Validity

In the context of risk adjustment for studying health care outcomes, *attributional validity* reflects the extent to which observers can identify the real causes of variations in outcomes across patients (e.g., the reason for differences in mortality rates). Donabedian (1980, 103) described this critical concept:

> When outcomes are used to make inferences about the quality of care, it is necessary first to establish that the outcomes can, in fact, be attributed to that care. We may call this the problem of "attribution," and its satisfactory solution may be said to confirm "attributional validity."

Attributional validity is critical when risk-adjusted outcomes information is used to motivate practice changes or monitor providers. For example,

comparisons of risk-adjusted outcomes with good attributional validity should provide rigorous insight about the causes of observed differences in the outcomes (i.e., whether differences can be attributed to the actions of clinicians or some other cause). However, attributional validity is difficult to measure. In the health policy arena, many assume that if a risk adjuster meets other validation standards (e.g., face validity, a certain level of predictive validity, or R^2), one can safely eliminate differences in patient risk factors as an explanation for differences in outcomes, such as mortality rates in performance profiles. However, one should exercise caution when attributing differences in risk-adjusted outcomes to various causes, even when using the "best" measure for risk.

Sometimes the data used for risk adjustment raise questions about attributional validity. One example involves the risk-adjusted mortality models produced by the University HealthSystem Consortium (UHC), an alliance of about 100 academic medical centers and nearly 200 affiliate hospitals formed in 1984 (Kozower et al. 2009). Using discharge abstract data and All Patient Refined Diagnosis-Related Groups (APR-DRGs), UHC produced risk-adjusted mortality models as benchmarks for quality assessment across its member institutions. Thoracic surgeons worried that problems with administrative data—specifically difficulties isolating postoperative complications from preoperative risk factors (see the discussion of present on admission [POA] indicators in Chapter 5)—might compromise the ability of the UHC risk-adjusted mortality rates to identify quality problems. Clinical data collected as part of the Society of Thoracic Surgeons (STS) cardiac database, created in 1986, do not have that problem (i.e., postoperative complications are clearly identified).

Kozower and colleagues (2009) compared hospital mortality predictions using UHC versus STS risk-adjusted mortality models for adult cardiac surgery patients and assessed how postoperative complications contributed to the performance of the two risk adjustment methods. As others have found, the R^2 and c-statistics for UHC's APR-DRG, administrative data–based risk adjustment method were superior to those of the STS clinical models (Kozower et al. 2009). The investigators showed, however, that the superior predictive ability of UHC's risk model came from its inclusion of postoperative complications. When the UHC model was evaluated on a subpopulation that specifically excluded patients with complications, its statistical performance metrics fell sharply, leading the researchers to conclude "[t]he current UHC risk-adjusted mortality models are not appropriate for preoperative quality assessment" (Kozower et al. 2009, 556). In other words, they questioned the attributional validity of benchmarking hospital quality using the UHC APR-DRG model, largely because of problems relating to administrative data.[3]

Assessing attributional validity is challenging. Often information about the ultimate target of interest (e.g., quality of care) is missing or of questionable

validity. The study in Massachusetts comparing hospital mortality performance rated by different risk adjustment methods, including a new model developed by UHC and its partner Premier (Shahian et al. 2010; see Chapter 2), provides good examples of the challenges encountered in assessing attributional validity. Nonetheless, especially in the current health policy environment—where payers might use risk-adjusted information to set provider reimbursement, for example—attributional validity is critical.

Examples of Validation Studies

Validity should be assessed within the context (outcome, timing, population, purpose) in which a risk adjuster will be used. Because researchers' access to data is restricted and options for evaluating validity are limited, creative approaches often are needed. The following sections describe two efforts to validate risk adjustment methods and risk-adjusted results. The first draws from work conducted by Yale University contractors to validate the mortality measure they developed for the Medicare Hospital Compare initiative (Krumholz et al. 2007); the latter comes from the NSQIP (see Chapter 8). These two examples describe potential strategies for validation studies but are not an exhaustive presentation of risk adjustment model validation methods.

Medicare Hospital Compare

As described in Chapter 1, Medicare's first foray into public reporting of hospital mortality rates in 1986 stumbled badly because of inadequate risk adjustment. The facility identified as having the most egregious mortality experience (i.e., the largest discrepancy between observed and expected mortality rates) in the country turned out to be a hospice, which cared for terminally ill individuals. The government's risk adjustment method failed to account adequately for patients' severity of illness when calculating their expected mortality rates. The government stopped reporting mortality rates early in the Clinton administration because of concerns that the administrative data–based risk adjustment methods failed to account sufficiently for health risks related to medical indigence and thus typically identified public hospitals ("safety net" facilities) as the worst performers.

In the early 2000s, however, CMS decided to try again, venturing into public reporting of hospital performance along a variety of measures. The new Medicare Hospital Compare website displays process-of-care measures for selected conditions (hospitals voluntarily submit data gathered through medical record review), volumes and payments for selected MS-DRGs, and mortality rates computed by the government using readily available administrative data. While earlier mortality reports focused largely on hospital-wide

death rates, the new initiative concentrated on death rates for specific conditions: acute myocardial infarction (AMI), heart failure (HF), and pneumonia. To develop and validate the condition-specific, risk-adjusted mortality measures, CMS contracted with researchers at Yale University led by cardiologist Harlan M. Krumholz; Sharon-Lise T. Normand from Harvard University was the lead statistician. The full report documenting the development and validation of the Hospital Compare risk adjusters for AMI and HF is available from the CMS website (Krumholz et al. 2007).

Despite the previous problems with risk adjusters derived from administrative data, CMS required the new models to use that data source because of their ready availability. However, the researchers used several strategies to enhance the clinical meaningfulness of their administrative data–based measures, including (Krumholz et al. 2007, 8–11)

- trying to differentiate complications from comorbidities (the data set used to produce the models did not yet include the POA flag; see Chapter 5);
- using a standard time window of observation (30 days post-admission) to avoid problems relating to different hospital discharge practices;
- including information from prior hospitalizations, such as previous coronary artery bypass graft surgery and diagnoses from previous admissions (restricting analyses to individuals with at least 12 months of fee-for-service enrollment prior to the index hospitalization); and
- grouping the 15,000 ICD-9-CM diagnosis codes into clinically coherent categories for modeling (they used the 189 diagnostic groups contained in the Hierarchical Condition Categories [HCCs]).

Krumholz and colleagues (2007) used two primary approaches to validate their models. The first approach involved statistical validation, primarily splitting the original sample in half (into derivation and validation data sets) and recalibrating the model variables using the validation data. They also performed additional statistical validations with other validation samples. The second, gold-standard approach took advantage of the existence of extensive information abstracted from medical records for AMI and HF patients, which could be linked to the patients' administrative data. For all validation efforts, the researchers produced indexes of predictive ability, discriminant ability, and overall fit. For the gold-standard analyses, they also compared the ability of the administrative data– and chart review–based methods to classify hospital performance. Analyses considered only Medicare beneficiaries aged 65 or older.

The researchers produced the derivation and validation data sets by randomly sorting patients, stratified by hospital, into equal halves (Krumholz et al. 2007). After performing the validation using the split-half data set, they

performed subsequent validations with individual one-year administrative data sets from several years subsequent to the initial data set year. Using the derivation half of the data, the researchers estimated two types of regression models: (1) a generalized linear model (GLM) linking predictor variables (i.e., potential risk factors) to the outcome (death within 30 days of admission) and (2) a hierarchical generalized linear model (HGLM) to account for the clustering of observations within hospitals (HGLM links risk factors to the outcome and a hospital-specific random effect; see Chapter 12 for discussion of the pros and cons of hierarchical models). HCC-based variables were used to represent relevant diagnoses, along with other candidate risk factors. Because the researchers intended to build a parsimonious model, they eliminated candidate predictor variables by using a backwards elimination procedure based on appropriate "exit" criteria and clinical judgment about the medical meaningfulness of predictors (see Chapter 10 for more information about model development in general).

For the statistical validation, Krumholz and collaborators (2007) used the remaining half of cases in the validation sample to validate the model developed on the derivation data set. When additional years of data became available, they reestimated the model coefficients using 100 percent samples of these additional annual databases. They recalibrated the predictor variable coefficients for these new samples, looking at consistency across time, and examined how frequently predictors were selected in different databases. After arriving at a final model, the researchers combined the derivation and validation samples and refit the model, using all the data from the year. They then used the parameter estimates from this final model to produce the hospital-specific measures. To do so, they calculated a standardized outcome for each hospital "by computing the ratio of the predicted to expected mean outcomes multiplied by the unadjusted national mean" (Krumholz et al. 2007, 20). They used bootstrapping procedures to compute confidence intervals around these hospital-specific performance measures.

The gold-standard validation linked administrative data to information generated through detailed reviews of medical records (Krumholz et al. 2007). On the basis of medical literature reviews and clinical judgment, the researchers identified candidate variables to include in the models they generated using the detailed clinical data sets. They used correlation coefficients to quantify the relationships between the outcomes estimated using risk factors from the clinical data and those estimated from administrative data. "A high value of the correlation coefficient provides evidence that the estimates based on administrative data are strongly linearly related to those based on the chart data. We also evaluate the graphical association between estimates of the two models" (Krumholz et al. 2007, 22).

For AMI cases, the gold-standard clinical data came from approximately 200,000 admissions abstracted during the Cooperative Cardiovascular

Project. Of these admissions, the researchers linked 181,032 cases to administrative data. The data included vital signs, cardiac symptoms, time since chest pain began, initial laboratory results, and electrocardiogram findings, among other factors. "While the administrative models explained about 10% of the observed variation and had accuracy of 69%, the chart model explained 24% of the variation and had accuracy of 77%" (Krumholz et al. 2007, AMI-46). Between the two data sources, the correlation coefficient of standardized mortality rates was 0.90 (Krumholz et al. 2007, AMI-51). The c-statistics for the HGLM derived from the administrative data and the GLM were 0.71 and 0.70, respectively.

For HF cases, the gold-standard data came from an initiative undertaken to examine quality of care across the 50 states, the District of Columbia, and Puerto Rico. Approximately 800 HF cases were identified from each state, and two clinical abstraction centers reviewed the records. The data set totaled approximately 46,700 cases once various exclusion criteria were applied to eliminate certain cases. The data included vital signs, cardiac symptoms, initial laboratory results, cardiac and noncardiac history, and left ventricular ejection fraction, among other factors. The final model contained 28 covariates; age had the largest effect on risk. "While the administrative mortality models explained about 10–12% of the observed variation and had accuracy of 69–71%, the chart model explained 21–22% of the variation and had accuracy of 75–78%" (Krumholz et al. 2007, HF-81). The correlation between the administrative and chart review models for 30-day mortality was 0.95.

NSQIP and Attributional Validity

As described in Chapter 8, the VA surgical outcomes initiative was designed to improve quality of care by giving surgeons information with sufficient clinical credibility to motivate self-examination and change. Although the VA investigators initially developed their risk adjustment models at the individual patient level, they aimed to generate their most important reports for each VA medical center performing surgery. VA surgeons participated actively in ensuring the face validity of the risk adjustment models constructed for individual patients (Khuri et al. 1995, 1997; Daley et al. 1997a, 1997b). In 1994, these risk adjustment methods became the foundation of an ongoing quality assessment activity that eventually became the American College of Surgeons' NSQIP. For the VA, however, tracking operative outcomes at its 123 hospitals nationwide remained its priority (Khuri et al. 1998; Daley, Henderson, and Khuri 2001). For this purpose, attributional validity was key: Did a valid causal link exist between poor risk-adjusted hospital performance and the quality of the hospital's operative care? Did hospitals with higher-than-expected operative mortality and morbidity rates provide lower-quality care than did hospitals with better-than-expected outcomes?

The VA investigators studied attributional validity in two ways: chart reviews and site visits. The chart review study examined 739 general, peripheral vascular, and orthopedic surgery records from the 44 VA hospitals participating in the national VA Surgical Risk Study (Gibbs et al. 2001). The researchers sampled cases for two levels of analysis: hospitals (facilities with higher- and lower-than-expected mortality and morbidity rates) and patients (individual patients with high and low predicted likelihood of mortality and morbidity who had died or developed complications). Twenty-one VA general surgeons and eight VA vascular surgeons reviewed the records using a structured implicit review instrument. Kappa analyses found fair to good agreement across the surgeons' judgments about quality (Gibbs et al. 2001, 190).

The chart review study produced mixed findings about the validity of the risk predictions (Gibbs et al. 2001). Patients from hospitals with higher- and lower-than-expected mortality and morbidity rates had similar overall quality of care. However, patient-level analyses found that patients with low predicted likelihoods of adverse outcomes who nonetheless experienced adverse outcomes had worse quality of care than did those with high predicted risks. The researchers speculated that chart reviews may be relatively insensitive to hospital-level quality problems; medical records may not contain clues about systemic quality difficulties within institutions. Thus, the investigators questioned the validity of the tool (chart reviews) they used to test the validity of risk-adjusted hospital findings as a quality indicator.

The second attributional validity study involved hospital site visits (Daley et al. 1997b; Young et al. 1997). Because many attributes of surgical practices within institutions cannot be evaluated through chart reviews, the researchers visited 20 outlier institutions: ten surgical services with the highest risk-adjusted mortality or morbidity rates and ten surgical services with the lowest mortality or morbidity rates. The visitors were blinded as to whether institutions were high or low outliers. The site visit teams, which consisted of a chief of surgery, a surgical ICU nurse specialist, and a study investigator, spent two days observing and interviewing staff at each surgical service. They examined the technology, equipment, and physical structure of the surgical service; the technical competence of the surgeons, anesthesiologists, nurses, and house staff; the service's relationship with the affiliated university's teaching programs; the service's relationships with all other patient care services in the hospital (e.g., internal medicine, radiology, laboratory); quality monitoring and improvement activities in the surgical service; communication and coordination among surgery, nursing, and anesthesia; and leadership in surgery and the institution.

The site visitors identified significant differences between the high and low outliers in several dimensions. Surgical services with better-than-expected

outcomes were more likely to have better technology and equipment (e.g., more up-to-date anesthesia equipment and monitoring equipment in the surgical ICU). These low-outlier institutions were more likely to standardize their approaches to routine patient care, such as by using practice guidelines or clinical pathways for routine operations. Thus, care was typically better coordinated at low-outlier hospitals. The site visitors, unaware of the outlier status of the surgical services they visited, rated overall quality of care higher among the ten low-outlier surgical services than among the ten high-outlier services. Site visit results strongly supported the attributional validity of hospital mortality and morbidity rates as an indicator of the quality of surgical care.

Reliability

An important indicator of the quality of a scientific measure is its ability to yield consistent, reproducible results. Statisticians call this characteristic *precision*, whereas social scientists, psychologists, and health services researchers know it as *reliability*. Both terms address this basic question: If this process were repeated, possibly by someone else but following identical rules, would the same results occur? Measures can be reapplied by the same observer or rater (to test intrarater reliability) or by different raters (to test interrater reliability). Agreement among different raters is a more rigorous test. As Feinstein (1987, 170) noted, reliability matters: "No matter how the observations are made and described, the data will have scientific quality if the results of the observational process can be consistently reproduced by the same or another observer." The following sections explore reliability as it pertains to risk adjustment, drawing heavily from Hughes and Ash (1997). These sections do not comprehensively review technical methods for assessing reliability, which are better described in methodological texts.

Reliability and Data Sources

As anticipated by Donabedian's quotation at the outset of this chapter, the accuracy or reliability of the underlying data carries immediate implications for risk adjustment methods. Chapters 5, 6, and 7 discuss reliability issues relating to the three main data sources (administrative databases, medical records, and patients). Here, I underscore several points.

The reliability of administrative data–based risk adjusters revolves around how accurately and completely information is reported from the medical record or other primary sources, such as how consistently diagnosis and procedure codes are assigned or race and ethnicity data are gathered. In studies using administrative data, returning to medical records to see if reabstractions yield identical results is virtually infeasible. The reliability of administrative

data-based risk adjusters is thus not tested directly but inferred from studies of coding reliability. Such reports rarely pertain to the specific database being used but instead relate to Medicare data in general or other administrative data sources (see Chapter 5). This imperfect approach to assessing reliability is a concession to the realities and costs of data collection.

For risk adjusters based on clinical data, reliability considerations are more complicated. When data are drawn automatically from an electronic source, such as a laboratory reporting system, one can assume this process is performed reliably (although questions might arise about how accurately a laboratory determines a clinical parameter). When data are manually abstracted (even from EHRs), reliability relates to the number of data elements, the complexity and subjectivity of judgments required to identify variables, the quality of the records (e.g., completeness, organization), and other factors (see Chapter 6). Studies using medical records should conduct reliability testing. Unreliable data may not present problems when variations occur randomly across patients or hospitals, but systematic variations may affect risk-adjusted findings. Data gathered on complex patients with numerous clinical findings may be less reliable than data on less complicated cases. Researchers should not use data until reliability reaches acceptable levels.

Risk adjusters based on patients' reports (e.g., reports of medical conditions or functional status) must accept some level of unreliability or subjectivity. Studies using patients' reports often trade data quality concerns for ease of administration or the inherent validity of patients' self-reports (i.e., as indicating patients' perspectives on a particular topic, such as their physical functioning or quality of life). The question is whether differences in data quality are nonrandomly distributed across patient types or various units of observation, such as hospitals or health plans. Data quality may vary nonrandomly, as the accuracy of self-reports may relate to education, income, literacy, and language—quantities distributed perhaps nonrandomly across hospitals, health plans, and other groupings of patients and providers. The reliability of patient-reported data is logistically complicated and costly to determine.

Sources of Variability

Variability may be inherent in the underlying data elements (e.g., laboratory tests, self-reported functional status), result from performing the risk adjustment, or be introduced by the persons applying the risk adjuster. Feinstein (1987) called these three sources of variation *input, procedure*, and *user* variability, respectively. These sources, discussed in the following paragraphs, may overlap.

When one is evaluating a biological variable, such as serum cortisol level, which varies systematically throughout the day, taking samples at a specified time could reduce diurnal variation. Similarly, repeated assessments of mental

function of persons with dementia may yield different scores day to day relating to mood, level of sedating medications, or sleep adequacy. These examples illustrate variation in the entity under evaluation, or *substrate*.

For risk adjusters, medical records present an analogous situation. For example, input variability pertains largely to the reproducibility of clinical information in the medical record. Once created, medical records (even EHRs) should not change. Nonetheless, several sources contribute to variability during the creation of records, including the setting and type of organization generating the record (e.g., tertiary teaching hospital, individual physician office); information management practices at the organization producing the record, including the interface of electronic and paper data sources, and other policies (e.g., how many diagnoses are retained); the structure of the medical record, including whether it contains templates and set response categories; and the types of illness under study (see Chapter 6). As noted previously, the copying and pasting of text in EHRs may perpetuate errors. In risk adjustment studies, these factors often vary across the units of observation, such as hospitals, physician practices, or health plans.

Whereas input variability relates to creating medical records, procedure variability concerns their use. Procedure variability arises in applying risk adjustment methods to medical records (i.e., extracting information from EHRs or charts and applying the formula or algorithm to generate the risk score). Procedure variability encompasses several components. First, risk factor information may be inherently unreliable. Although training, diligence, and close supervision of data abstractors may lessen this unreliability, it cannot be eliminated without automated, electronic algorithms that collect clear-cut, objective information. Automated extraction of risk factors from narrative sections of EHRs (e.g., clinicians' notes) is susceptible to the vagaries of coding schemes, intensity of documentation, and variations in language use. Templates for data gathering and preset response categories might help if the information sought is clearly specified and language is universally used—a difficult standard to meet. Thus, even when information is extracted from EHRs, risk factors fall into a hierarchy of reproducibility. Numerical values are abstracted most reliably, whereas data derived from narrative texts are generally least reliable.

A second source of procedure variability is the replicability of the rating scheme—whether users can follow abstraction and rating procedures as readily as the original developers can. Again, computerized scoring algorithms largely eliminate this concern. When manual methods are used, vague, ambiguous, or incomplete instructions compromise consistent scoring.

A third concern is fragility—the extent to which risk adjusters are susceptible to a few influential risk factor values. Fragile risk adjusters are strongly influenced by rare but idiosyncratic findings, particularly those that are

unreliably detected. Evaluating the fragility of a code-based risk adjuster involves, for example, examining how ratings change following modifications to the choice and ordering of diagnosis codes. Methods that rate cases on the basis of a single diagnosis code are less reliable than those that rely on multiple codes; in the former instance, a single coding inconsistency could significantly affect the risk rating. Clinically based methods raise similar concerns about fragility, especially when they rely on a single data element (e.g., single temperature reading or laboratory value) to rate cases. Some clinically detailed risk adjusters require two or more values of a given parameter (e.g., two or more highly elevated temperatures) to assign a risk factor (e.g., high fever).

With the possible exception of some laboratory and diagnostic test results, all risk factors—from diagnosis codes to ratings of self-perceived health—carry some component of human interpretation or judgment. Persons using risk adjustment methods are motivated not only by their personal characteristics but also by external attributes, such as their employer, their perception of their personal and professional roles, how they are trained and monitored, stress and time pressures, work environment (e.g., ventilation, lighting), the nature of their working relationships, and organizational culture. For example, coding practices vary among hospitals and possibly even geographic regions (Romano et al. 1997; Welch et al. 2011). Hospitals may aggressively pursue comprehensive coding if they are paid for treating sicker patients or worry about external publication of risk-adjusted report cards. On the other hand, the threat of being fined for overcoding is a disincentive to coding diagnoses that may be present but would not be easy to defend in an audit.

The training and professional background of the person reviewing the record also are critically important. As noted in Chapter 5, diagnosis and procedure coding using ICD-9-CM (and soon ICD-10-CM and ICD-10-PCS) requires extensive training. Although hospitals usually employ individuals with technical coding degrees to produce their discharge abstracts, coders in other settings, such as physicians' offices, often have less training. The background of the reviewers can compromise reliability in unexpected ways. My colleagues and I once hired a research staff to abstract medical records with MedisGroups, a clinically detailed risk adjustment method based on key clinical findings (KCFs) abstracted from medical records (Iezzoni et al. 1992a). The reviewers—registered nurses with additional postgraduate training—were uncomfortable following strict data collection protocols established by the developers of MedisGroups to achieve high reliability. Although the guidelines often specified the exact words needed to establish the presence of a KCF, the reviewers wanted to use their extensive knowledge to interpret findings or reflect their own clinical sense of a case. Here, independent judgment, which is highly valued in clinical care, led to unreliable data collection.

Assessing Reliability

Reliability assessment aims to determine the amount of agreement resulting from repeated application of the risk adjuster. The simplest measure is the percentage agreement between reviewers or raters. MedisGroups used this strategy to rate reviewer accuracy. During training, new MedisGroups chart reviewers compared their KCFs with those identified by expert instructors; 95 percent agreement was required before the trainee qualified as a MedisGroups reviewer.

High percentages of exact agreement may occur, however, even without skilled reviewers or reliable review instruments. For example, rater A examines 100 cases, classifying 90 as high risk and 10 as low risk. Rater B labels patients randomly, classifying 80 as high risk and 20 as low risk. The expected percentage agreement (chance agreement) between raters A and B is 74 percent: Raters A and B are expected to agree on 80 percent of the 90 high-risk cases ($n = 72$) and 20 percent of the 10 low-risk cases ($n = 2$). Suppose the actual observed agreement turns out to be 75 percent, which is nearly equal to the 74 percent chance agreement. If rater C simply called all 100 cases high risk, raters A and C would have 90 percent agreement on the high-risk cases.

Cohen's Kappa

Kappa is the most common measure of the reliability of instruments with categorical and ordinal scales (Cohen 1960). Kappa can take several forms. The simplest form measures how much the level of agreement between two observers exceeds the amount of agreement expected by chance alone, computed as

$$K = \frac{P_o - P_c}{1 - P_c},$$

where P_o is the observed agreement and P_c is the agreement that would have occurred by chance. In the previous example of raters A and B using a two-level scale, if the overall agreement rate were 75 percent, kappa would be

$$\frac{0.75 - 0.74}{1.00 - 0.74} = 0.04.$$

How "good" is a kappa score of 0.04? In their seminal article, Landis and Koch (1977) assessed kappa values as follows: kappa <0.0, poor agreement; kappa 0.0–0.20, slight agreement; kappa 0.21–0.4, fair agreement; kappa 0.41–0.6, moderate agreement; kappa 0.61–0.8, substantial agreement; and kappa 0.81–1.0, almost perfect agreement.

Although kappa is superior to percentage agreement, it becomes particularly unstable when the prevalence of the target outcome is small. For example, suppose two raters review the same 100 cases and each finds two problem cases. Three scenarios are possible (Exhibit 9.2): They find the same two problem cases; they agree on just one case; or each finds two different cases. Percentages of exact agreement for these three scenarios are 100 percent, 98 percent, and 96 percent, respectively; kappas are 1.00, 0.49, and -0.01. Disagreement on one or two cases—when the prevalence of the problem is small—thus can cause kappa to vary over the entire range of possible values.

The simple form of kappa applies to assessment of agreement between two raters. In cases of more than two raters, kappa scores can be computed for each rater pair and then averaged.

Weighted Kappa

In the basic form of kappa, all disagreements among raters are rated the same regardless of the magnitude of disagreement. Difficulties arise when raters have three or more choices on an ordinal scale, such as an edema (or swelling) score that ranges from 1+ to 4+. The basic formulation of kappa can apply here but could obscure important differences in reliability. For example, suppose two ordinal scales yield possible scores from 1 to 4; both produce 70 percent agreement between two raters. For one scale, all disagreements are within one point. For the other, disagreements are frequently two or three points apart. In the basic form of kappa, both scales achieve the same kappa value, although the sizes of disagreements for the two scales differ widely.

Landis and Koch (1977) proposed ways to weight kappa to give partial credit for small discrepancies or increase the penalty as the magnitude of the discrepancy among ratings increases. For example, when exact agreement is given a value of 1, ratings for an individual case within one rank might be given a value of 1/2, those within two ranks might be given a value of 1/4, and those more than two ranks apart might be given a value of 0. Another approach is to base the score on the square of the amount of disagreement, which causes the

EXHIBIT 9.2 Agreement of Ratings Between Two Reviewers

Scenario 1 (Kappa = 1; percentage agreement = 100%)

		Rater 1	
		No	Yes
Rater 2	No	98	0
	Yes	0	2

Scenario 2 (Kappa = 0.49; percentage agreement = 98%)

		Rater 1	
		No	Yes
Rater 2	No	97	1
	Yes	1	1

Scenario 3 (Kappa = −0.01; percentage agreement = 96%)

		Rater 1	
		No	Yes
Rater 2	No	96	2
	Yes	2	0

penalty for disagreement to increase geometrically as the discrepancy among ratings increases. Because this form of weighted kappa depends on the squared value of the difference in ratings, it is an analysis of variance (ANOVA) procedure, which provides opportunities for other statistical manipulations.

Other Measures of Reliability

The *intraclass correlation coefficient* (ICC), or *interrater reliability coefficient* (RI), is based on one-way ANOVA and may be used in one of several forms to assess the degree of agreement among two or more raters. Because it is a proportion, the ICC ranges from 0 (complete disagreement) to 1 (complete agreement). Shrout and Fleiss (1979) developed several types of calculations for ICC, each a ratio of variance estimates. In each, the variance of ratings for different cases—the between-mean square (BMS)—is divided by the sum of BMS and other variance components that depend on the purpose of the particular analysis.[4] Because the ICC is based on one-way ANOVA, it treats risk adjusters as interval scales. For risk adjustment methods that produce scores using ordinal scales, use of the ICC could lead to misleading reliability results. Shrout and Fleiss (1979) provide an excellent discussion of considerations regarding selection of the appropriate form of the ICC.

Kendall's tau (sometimes called *gamma*) is useful for evaluating ordinal scales because it measures the degree to which two observers rank cases in similar order. Kendall's tau is calculated by taking all possible pairs of cases and noting how often the pairs have concordant and discordant rankings (i.e., are ranked in the same order versus the opposite order by both observers; ties are not counted). Specifically, tau equals the number of concordant rankings minus the number of discordant rankings divided by the total number of pairs. The score can range from −1.0 (implying total disagreement) to 1.0 (total agreement). Several variations of tau (tau_a, tau_b, and tau_c) are useful if ties are present in rankings. Gibbons (1993) further describes these measures and provides simple examples.

Other Issues in Assessing Reliability

Different risk adjusters quantify risk along ordinal or interval scales, a distinction that has important implications for reliability measurement. An interval scale provides a continuum of values; it suggests that a given increase in the scale always has the same implication for increased risk. On an interval scale, for example, the difference between scores of 5 and 10 conveys the same increased risk as the difference between scores of 25 and 30. In contrast, an ordinal scale conveys only a ranking; a higher score simply indicates greater risk. For example, on an ordinal scale, patients with scores of 3 face higher risks than those with scores of 2, but the difference between

scores of 2 and 3 is not necessarily equivalent to the difference between scores of 3 and 4.

The reliability of a measure is to some extent an artifact. For example, suppose there are two raters, and each rates 100 persons on a three-point scale (the results are shown in Exhibit 9.3). The kappa value is (60 – 34) ÷ (100 – 34) = 0.39, which is fair by the standards of Landis and Koch (1977). However, one can boost kappa to 0.52 (moderate) by collapsing the observations to a two-point scale (see Exhibit 9.4). The perceived improvement is somewhat real, as many misclassifications relate to distinctions between the scores 1 and 2. While one can often improve reliability by collapsing scores, reducing the number of categories may also cause kappa to fall, as shown in Exhibit 9.5, when a different two-point scale is used. Here, kappa drops to 0.38. These examples highlight two important lessons: (1) Do not rely too heavily on precise values of kappa, and (2) do not compare the kappas of risk adjusters that distribute cases into categories very differently.

EXHIBIT 9.3 Agreement Between Two Raters Using a Three-Point Scale

		Scores of Rater 1			
		1	2	3	All
Scores of Rater 2	1	25	10	5	40
	2	10	15	5	30
	3	5	5	20	30
	All	40	30	30	100

Kappa = 0.39; percentage agreement = 60%

EXHIBIT 9.4 Agreement Between Two Raters, Collapsing a Three-Point Scale to a Two-Point Scale: Case 1

		Scores of Rater 1		
		<3	3	All
Scores of Rater 2	<3	60	10	70
	3	10	20	30
	All	70	30	100

Kappa = 0.52; percentage agreement = 80%

EXHIBIT 9.5 Agreement Between Two Raters, Collapsing a Three-Point Scale to a Two-Point Scale: Case 2

		Scores of Rater 1		
		1	>1	All
Scores of Rater 2	1	25	15	40
	>1	15	45	60
	All	40	60	100

Kappa = 0.38; percentage agreement = 70%

Although a risk adjuster's reliability is important, measuring reliability presents considerable challenges. A single measure of interrater reliability is generally insufficient. In any case, the ultimate test of reliability for any risk adjuster is not how well it performs under laboratory conditions but rather its performance in the real world. Real-world dynamics, such as financial incentives and concerns about public disclosure, may be the most powerful forces that compromise reliability.

Notes

1. Some controversy exists about distinctions between clinimetrics, a field pioneered by Feinstein (1987) that focuses on the quality of clinical measurements, and psychometrics (Streiner 2003; Fava and Belaise 2005).
2. Sinuff and colleagues (2006) searched MEDLINE (1966–2005), CINAHL (1982–2005), Ovid Healthstar (1975–2004), EMBASE (1980–2005), SciSearch (1980–2005), PsychLit (1985–2004), the Cochrane Library (Issue 1, 2005), PubMed "related articles," personal files, abstract proceedings, and reference lists. They looked for "all studies that compared physician predictions of ICU or hospital survival of critically ill adults to an objective scoring system, computer model, or prediction rule. [They] excluded studies if they focused exclusively on the development or economic evaluation of a scoring system, computer model, or prediction rule" (Sinuff et al. 2006).
3. Kozower and colleagues (2009, 556) did not believe that POA flags would remedy this problem: "Although POA coding will make a substantial improvement in the measurement of baseline risk, it will not solve all of the problems with the UHC or other models relying

on administrative data. For example, diagnosis codes for situations such as an intraaortic balloon pump inserted at a different hospital before transferring a patient will not be accounted for in the APR-DRG system. The importance of this will vary significantly depending on the referral and transfer patterns of a hospital."

4. In general, a mean square is calculated from a sum of squares by dividing the sum of squares by $(n - p)$, where n is the sample size and p is the number of estimated parameters. The most familiar example of this calculation is its use for determining the variance of a set of data. The sum of squares total, or SST, is $\sigma_i(Y_i - Y)^2$. The mean square, which in this simple case is used as an estimate of the population variance, is SST $\div (n - 1)$. Here, n is the sample size and 1 is subtracted because just one parameter, the population mean, has been estimated (using Y) from the data.

CHAPTER

10

EMPIRICALLY EVALUATING RISK ADJUSTMENT MODELS

Michael Shwartz and Arlene S. Ash

The first question asked about a risk adjustment method often is, "How good is it?" Questioners generally want a simple answer: a number that enables them to compare one risk adjuster's performance to another's. They also may want to know its *predictive validity:* how well the risk adjuster accounts for actual differences in patients' risks for particular outcomes. As noted throughout this book, a summary indicator of statistical performance does not reveal whether a risk adjustment method is valid for its intended purpose. Other important dimensions of validity, including clinical credibility and reliability of the underlying data sources (see chapters 8 and 9), determine whether persons reviewing risk-adjusted outcomes data believe and act on the information.

Statistical performance is evaluated by calculating one or more summary measures of the level of concordance between the risk-adjusted predictions and actual outcomes of individual patients in a population. It is usually possible, and generally desirable, to compare models—even those derived in different ways—by "using the same yardstick." Many excellent textbooks discuss the principles and practices of empirical modeling. Two books of special relevance to health services researchers are Harrell (2001) and a less technical exposition inspired by Harrell (Steyerberg 2008).

In this chapter, we examine special concerns about the use of multivariable models to adjust for differences in risk in assessment of health care outcomes and discuss quantitative measures of the validity of models predicting either continuous outcomes (e.g., lengths of stay, costs of care) or dichotomous outcomes (e.g., death, the presence of a complication). Throughout, we draw examples from risk adjustment methods that have been well described in the published literature and in our own work. First, we consider how to convert risk scores into expected (or predicted) outcomes (*PRED*s).

Translating Risk Scores into Predicted Outcomes

Risk adjustment requires estimating expected outcomes for individuals on the basis of their risk factors. For each person, let Y be the outcome (e.g.,

hospital charges) and *PRED* be the expected value of *Y* for that person. For a group of people, we compute the expected value of the sum of group members' *Y*s (e.g., total charges for all patients in a hospital) by summing their *PRED*s. Equivalently, the mean of their *PRED*s is the expected mean for their *Y*s (expected charge per patient). If *Y* is a 0/1 outcome, such as 1 = died versus 0 = lived, each *PRED* is the probability that a given individual in the group will die. The sum of group members' *PRED*s is the expected number of deaths in the group; the mean of their *PRED*s is their expected death rate.

Even when risk scores are derived explicitly to predict the outcome of interest, analysts may want to modify the scores before using them as *PRED*s, depending on the purpose of the analysis. Suppose that 10 percent of 10,000 acute myocardial infarction (AMI) patients hospitalized in California died but that the expected death rate, based on a risk adjustment model, was 12 percent. To the extent that the model's predictions provide an appropriate standard, California outcomes are relatively good, perhaps because overall quality of care is better there. Other factors also may cause observed and predicted outcomes to diverge. Identifying the reasons for a divergence is difficult and controversial. For example, in-hospital risk-adjusted coronary artery bypass grafting (CABG) mortality rates in New York State declined from 4.2 percent in 1989, when the state began publicly disseminating such data, to 2.5 percent in 1992 (Hannan et al. 1995). The decrease in risk-adjusted rates was caused by both a decline in actual mortality (from 3.5 to 2.8 percent) and an increase in expected mortality (from 2.6 to 3.5 percent). These declines have been attributed to improved performance (Hannan et al. 1995; Chassin, Hannan, and DeBuono 1996), but non-quality-related factors also may be contributing. As reviewed by Epstein (2006), these factors include increased coding of comorbidities (which increases expected CABG mortality rates relative to observed rates and thus reduces the risk-adjusted rate), use of additional or different procedures (so that some cases are not classified as "isolated CABG" and are thus excluded from risk-adjusted CABG rate calculations), and avoidance of high-risk patients. To the extent that coding intensity has increased or the population of cases included in the analysis changes over time or differs across settings, comparisons to historical or external standards can be misleading. Even if models developed for one population predict incorrect averages for new populations, the relative risks encoded in the predictions (e.g., whether one patient is 50 percent higher risk than another) usually generalize well.

A simple and effective way to tailor a model's predictions to a new population is to rescale the model by multiplying each prediction by a constant, *K*:

$$K = \frac{\textit{new population average}}{\textit{predicated population average}}.$$

Rescaling forces the expected outcome to exactly equal the actual outcome for the new population's data. In the California AMI mortality example earlier, K equals 10/12. Rescaling is the most straightforward way to calibrate a model to new data. More sophisticated calibration not only helps produce correct mean values but also fine-tunes models to specific applications. Predictions that have been rescaled or more generally calibrated to a new population no longer provide an external standard for that population, but they remain useful for examining risk-adjusted differences in outcomes across policy-relevant subgroups. For example, in the California AMI scenario one could compare outcomes for Medicaid versus private-pay patients or among patients of different providers or of different health care delivery systems.

Sometimes risk scores are calibrated so that their average equals one for a benchmark population. Such scores must be calibrated to predict outcomes. For example, Diagnostic Cost Group (DCG) prospective relative risk scores (RRSs) estimate next year's expected total health care costs as a multiple of average costs (Ash et al. 1989; Pope et al. 2004). RRSs are calculated from demographic data and a vector of health status information. For example, an RRS of 1.2 indicates that expected resource consumption is 20 percent above average. As does any score developed to be proportional to the outcome of interest, relative risk scores need only to be rescaled.

More flexible calibration is required when risk scores are systematically associated with the outcome but not necessarily proportional to it. One simple calibration method is "bucketing." Analysts rank cases in the order of their risk scores and then assign cut points to define categories, called "buckets" or "bins."[1] No single bucket should contain most of the cases, and each should contain enough cases for calculation of a stable average.[2] The prediction for each case is the average of the actual outcomes for all cases in the same bucket.[3] Even when higher-scoring buckets truly have higher-risk cases, the cases in them can have lower average outcomes due to sampling variation. Indeed, such results are likely when some buckets contain few cases and when the outcome's standard deviation is large compared to its mean. When millions of cases are available for analysis, many buckets, such as one for each percentile of the predicted value, can be formed, each containing enough cases for calculation of stable averages. To avoid assigning lower *PRED*s to cases with higher risk scores—for example, when the highest-scored bucket has lower average risk than the second-highest-scored bucket—analysts can merge the two buckets, or a smooth, nondecreasing function can be fit to the data instead of "raw" bucket averages to predict risk. Bucketing approaches are particularly useful when risk scores are calibrated to one outcome (e.g., cost) but used to predict a related outcome (e.g., length of stay). Large data sets permit more refined recalibrations, such as regressing the outcome in the target population on the risk score (or indicator/dummy variables for "score buckets") and other factors (such as demographic characteristics).[4]

If a model is used to generate the *PRED*s, it usually consists of equations that specify coefficients for each explanatory variable (i.e., risk factor). These equations are then applied to new data to calculate a new risk score for each case.[5] In model development, the values of both explanatory and outcome variables for each case must be known. However, only the explanatory variable values for each case are required for application of the model because a model converts inputs (the values of the explanatory variables) into outputs (expected outcomes) using only the relationships identified in the development data set. Translating the risk score into *PRED*s may still require calibration involving judgment. For example, APACHE III scores *per se* do not estimate the probability of death, but by regressing an indicator for death on the scores, we can translate scores into predictions. With continuous scores, plots of the mean outcome versus the mean score within narrow, equal-width score categories can be examined to see if the relationship is well modeled by a straight line or if a more complicated functional relationship should be used for translation. A slightly more sophisticated technique uses a smoothed (nonparametric) regression algorithm to explore this relationship.[6] Another way to translate risk scores into *PRED*s is through bucketing (e.g., placing all cases with a risk score of 0 in one bucket, all cases with a risk score of 1 to 5 in a second bucket, all cases with a risk score of 6 to 10 in a third bucket, and so on), as discussed earlier.

The combined effect of a person's diagnoses on an outcome of interest has been especially well studied by researchers attempting to predict total health care utilization during a fixed period. For example, ACGs (originally Ambulatory Care Groups, now Adjusted Clinical Groups) were developed originally to estimate the resources needed annually to treat medical problems in ambulatory care settings (Weiner et al. 1991; Starfield et al. 1991). The ACG algorithm first assigns each **diagnosis** to one of 32 ADGs (originally Ambulatory Diagnosis Groups, now Aggregated Diagnostic Groups), such as "time-limited: major" or "likely to recur: discrete." Then each **person** is classified into one ACG on the basis of age, sex, presence of specified ADGs, and number of individual ADGs—for example, "acute minor conditions(s) only, age less than two" or "four or five different ADGs, age 17–44." Thus, each person is ultimately assigned to one of approximately 106 mutually exclusive ACG categories. Analysts must then produce scores to predict specific outcomes, such as number of visits, ambulatory charges, or total charges for the current year. One way to do so is to regress the outcome on a series of diagnosis-based dummy variables, age, and sex; another is to regress the outcome on dummy variables that identify the unique ACG categories. The latter method is a categorical (non-ordinal) equivalent of letting the *PRED* for each case equal the average value of the outcomes for all cases in the ACG category. Today's ACG system includes additional predictive modeling structures that use diagnostic data to predict a range of patient outcomes (www.acg.jhsph.edu).

DCGs also consider age, sex, and diagnoses (i.e., ICD-9-CM or ICD-10 codes) in predicting outcomes, such as next year's costs for Medicare beneficiaries. While an early version of DCGs used discharge diagnoses from hospitalizations alone (Ash et al. 1989), the DCG Hierarchical Condition Category (DCG/HCC) models developed in the 1990s (Ellis et al. 1996) make predictions on the basis of demographics (age and sex) and the presence of up to 184 condition categories based on diagnoses identified during all hospitalizations and ambulatory encounters with clinicians (Pope et al. 2004). Various features of the DCG models, including their hierarchies, were designed to make predictions that are minimally affected by variations in coding practice and that limit the potential effects of "code creep" (see Chapter 5) on calculation of payments to capitated health plans.[7]

Medicare uses versions of these DCG models to calculate payments to Medicare Advantage plans. Medicare models do not use the full complement of HCCs; instead they use a subset of HCCs. For example, the model name "CMS-HCC 70" indicates that 70 condition categories are allowed to contribute to increased plan payments. Starting in 2012, the Centers for Medicare & Medicaid Services (CMS) is raising the number of HCCs recognized in its payment calculations to 87 (CMS-HCC 87). Non-CMS-HCC models—referred to variously in the literature as DCG, DxCG, DCG/HCC, and DxCG/HCC models—are maintained and updated by Verisk Health, Inc. (VH, www.veriskhealth.com). VH-HCC models first construct a clinical profile for each person by considering which of 394 condition categories are present during some period (usually one year). Then, recognizing that multiple ICD-9-CM codes and even several condition categories could relate to a single medical problem, the models array the condition categories in hierarchies. Within a hierarchy, the presence of a more serious (higher-cost) diagnosis usually causes lower-cost, related diagnoses to be ignored. For example, codes for metastatic malignancy override all site-specific cancer codes, while all cancer codes dominate benign neoplasm codes.[8] The *PRED* for cost is derived by regressing cost on markers (dummy variables) for patients' demographics and HCCs and further tailored (using second-stage regressions) to ensure good fit for persons of diverse risk levels within fairly narrow age/sex bands. Verisk Health offers multiple DCG models that predict different cost, utilization, and quality outcomes from various data sources.

Model Performance in Development and Validation Data

It is important to distinguish how a model performs using the data from which it was developed (called the "fitting" or "development" data) versus how well it performs when applied to new data (called "validation" or "confirmatory" data) (see Chapter 9). The more variables a model includes, the greater the

importance of computing measures of performance using validation data. For example, to translate a risk score (RS) derived from a risk adjustment method into a predicted cost, analysts could calculate *PRED*s for a continuous outcome by fitting a simple regression form, such as:

$$\textit{Expected value}\ (Y_i) = E(Y_i) = a + bRS_i, \text{ which leads to } PRED_i = \hat{a} + \hat{b}RS_i.$$

An unknown variable that has been estimated from data is indicated by the "hat" symbol: ^; for example, an estimate of b is written as $\hat{b}$ and called "bee hat."

With only two coefficients to fit to a large data set, overfitting is not a major concern. However, in many situations, analysts use the same moderately sized data set to specify the model (i.e., to determine which of many risk factors to include and the form in which to include them) and to estimate coefficients for each included variable. This approach can lead to *overfitting*—that is, the variables identified and parameters estimated reflect, in addition to generalizable relationships between predictors and outcomes, the idiosyncrasies of the development data set. Complicated models with many risk factors that have been (over)fit to one data set may predict outcomes far less well when applied to new data.

When outcomes are approximately normally distributed, 30 cases per predictor variable (i.e., risk factor) are generally considered enough to establish reliable coefficients. For highly skewed outcomes, however, hundreds or even thousands of cases may be needed. Health care costs in the United States are one such highly skewed outcome. In 2006, the 50 percent of the noninstitutionalized US civilian population with the lowest costs accounted for just 3 percent of total health care expenditures. The top decile accounted for 65 percent of spending, the top 5 percent accounted for 49 percent, and the most expensive 1 percent accounted for 21 percent (Cohen and Yu 2010).

Especially when models include indicators for rare, expensive events, generic guidelines, such as no more than one predictor for each 30 cases when predicting continuous outcomes and no more than one predictor per 18 events when predicting dichotomous outcomes, are inadequate.[9] Ideally, even models built on large databases are validated on at least one other data set.

A practical impediment for model developers is lack of data for validation. As described in Chapter 9 and later in this chapter, analysts are able to estimate what the model's performance would be if it were applied to a validation database by creatively using a single development data set. On the other hand, when applying a model developed elsewhere, users are advised to examine validation statistics, such as those discussed later in this chapter, to see how well the intended model fits the data and purposes of a current project.

Measuring Model Performance for Continuous Outcomes

When asking how well a risk adjuster predicts a continuous outcome such as costs, most people expect to hear about the R^2 values. Although several definitions of R^2 exist, the most common is:

$$R^2 = 1 - \left[\frac{\sum_i (Y_i - PRED_i)^2}{\sum_i (Y_i - \overline{Y})^2} \right], \text{ where } \overline{Y} \text{ is the average of the } Y_i\text{s.}$$

In the numerator of the term subtracted from 1—that is, $\sum_i (Y_i - PRED_i)^2$—each $Y_i - PRED_i$ is called an *i^{th} residual, deviation,* or *error,* and the numerator itself is called the sum of squared errors (*SSE*). The denominator in the expression—that is, $\sum_i (Y_i - \overline{Y})^2$—is called the *sum of squares total* (*SST*). *SST* is determined by the data alone and not by the model; it measures the variability of the outcome Y in the data, while *SSE* measures the variability in Y that is not captured by the particular predictive model used to calculate the *PRED*s.

Let X_{ij} be the value of risk factor j for patient i. An ordinary least squares (OLS) algorithm finds the numbers $\hat{a}$ and $\hat{b}$ in the following equation that result in the smallest possible *SSE:*

$$PRED_i = \hat{Y}_i = \hat{a} + \sum_j \hat{b}_j X_{ij}.$$

Minimizing *SSE* is equivalent to maximizing R^2, so R^2 is a particularly appropriate measure of performance for models derived using OLS.

Even if predicted values are developed some other way, such as by using the binary splitting algorithm CART,[10] R^2 still can be computed and is a useful summary measure of performance. Examining the same performance measure is often desirable when comparing models constructed with different algorithms, even if these other approaches do not seek to maximize R^2. To easily calculate R^2 for a set of $(Y_i, PRED_i)$ pairs, first calculate r, the Pearson correlation of Y and *PRED*, and then square it.

R^2 is often described as the fraction of total variability in the outcome (dependent variable) explained by or attributed to differences in risk among cases included in the model. Sometimes analysts multiply R^2 by 100 to express it as the percentage of variation explained by the model. For example, Sibley and colleagues (2010) found that demographic information plus ACGs explained 33 percent of the variation in number of visits to a family physician over a two-year period and 21 percent of the variation in specialist physician visits. Adding survey-assessed measures of health status to the administrative data–based models further increased R^2 by less than 2 percent.

Most investigators routinely report R^2 to indicate how well risk adjustment models explain continuous outcomes, such as health care expenditures

and lengths of stay. For example, the Society of Actuaries (SOA) compared the performance of different risk adjustment methods based on different types of administrative data for predicting both this year's and next year's medical costs for a commercially insured population (Winkelman and Mehmud 2007). Of the systems SOA examined, we consider seven that used either diagnostic (indicated by Dx) or pharmacy (indicated by Rx) codes:

1. ACG, Version 7.1 (Dx): This method defines mutually exclusive groups according to morbidity-based ACG categories; selected high-impact, disease-specific Expanded Diagnosis Clusters (EDCs); and diagnostic indicators of the likelihood of future hospitalizations and of being medically frail (Winkelman and Mehmud 2007).
2. Chronic Illness and Disability Payment System (CDPS), Version 2.5 (Dx): This model, originally developed for use with Medicaid populations, assigns individuals to one of 16 age/sex cells and one or more of 67 medical conditions.
3. Clinical Risk Groups, Version 1.4 (Dx): On the basis of clinical criteria, a case is assigned to one of about 1,100 unique groups (the specific number depends on whether analyses are performed prospectively or retrospectively).
4. DxCG DCG, RiskSmart Version 2.1.1 (Dx): Diagnosis codes are grouped into 781 clinically homogeneous groups (DxGroups), which are then mapped into 184 HCCs. Each case is also assigned to one of 32 age/sex categories (Winkelman and Mehmud 2007).
5. DxCG RxGroups, Risk Smart Version 2.1.1 (Rx): National Drug Codes (NDCs) are classified into 164 mutually exclusive categories (called RxGroups) on the basis of each drug's therapeutic indication. Each patient is also assigned to one of 32 age/sex categories (Winkelman and Mehmud 2007).
6. Ingenix Pharmacy Risk Groups (PRGs), Version 5.3 (Rx): PRGs are based on the patient's mix of pharmacy prescriptions and how a drug relates to other drugs prescribed for the patient. Each NDC is mapped to one of 107 PRGs. Each case is characterized by age, sex, and PRGs (Winkelman and Mehmud 2007).
7. Medicaid Rx (Rx): Each individual is assigned to one or more of 45 medical condition categories on the basis of his or her prescription drug use and to one of 11 age/sex categories (Winkelman and Mehmud 2007).

The SOA's database, obtained from MedStat Marketscan, contained information on 620,000 members who were continuously enrolled for two years (2003 and 2004) in a comprehensive major medical insurance program. We consider R^2 values from analyses in which models fit to 2003 data were

used to predict 2004 claims (i.e., prospective models) and 2003 claims (concurrent models). Each model was evaluated twice: once using the risk weight for each age/sex and condition category supplied by the vendor of the particular risk adjustment method and once after recalibrating the weights to fit the SOA database.

Exhibit 10.1 shows the R^2s for predicting raw (untransformed) total costs and "top-coded" total costs, using $100,000 and $250,000 as thresholds for top-coding. As discussed later in this chapter, when costs are top-coded, any costs above the top-coding threshold are set at the threshold value. R^2s were higher, in some cases substantially so, when predicting top-coded costs; predicting less-skewed outcomes is easier. However, top-coded models do not account for all the dollars (i.e., the mean prediction for top-coded models equals the mean of top-coded costs, which is smaller than actual mean costs). Thus, the increase in R^2 associated with changing from models predicting raw costs to models predicting top-coded costs is expected. Of greater importance is how different models compare when predicting the same outcome (ideally, as in this case, on the same data set).

In contrast to the top-coding situation, the change in R^2 resulting from recalibration is interesting because analysts must choose between using risk scores from the vendors' algorithm and producing their own recalibrated coefficients or weights. In the SOA study, CDPS and Medicaid Rx achieved much higher R^2s when the weights were recalibrated. These methods used Medicaid data to produce their vendor-supplied weights while the SOA data represented a commercially insured population, so these differences are not surprising. The other methods achieved only modestly higher R^2s after recalibration, which should reassure researchers wishing to use methods "off the

EXHIBIT 10.1 Model Performance When Predicting Medical Costs from 2003 Diagnostic or Pharmacy Data

		Performance Measure: R-squared					Performance Measure: MAPE as a Percentage of Average PMPY Cost*
		Prospective (Predicting 2004)				Concurrent (Predicting 2003)	Prospective (Predicting 2004)
		Vendor-Supplied Weights			Recalibrated	Vendor-Supplied Weights	Vendor-Supplied Weight
		Topcoded at	Topcoded at	No	Topcoded at	Topcoded at	Topcoded at
Model	Inputs	$100,000	$250,000	Topcoding	$250,000	$250,000	$250,000
ACG	Diag	20.8	19.2	16.2	19.6	29.7	89.9
CDPS	Diag	17.6	14.9	12.4	17.7	32.9	95.3
CRG	Diag	19.3	17.5	14.9	NA	43.3	90.9
DCG	Diag	22.3	20.6	17.4	21.3	51.8	87.5
RxGr	Rx	23.8	20.4	16.8	20.5	NA	85.3
PRG	Rx	25.0	20.5	17.2	21.2	NA	85.8
MedRx	Rx	19.3	15.8	12.9	17.7	28.1	89.6

Diag: ICD-9 Diagnosis Codes

Rx: Pharmacy NDC Codes

*MAPE = mean absolute prediction error; PMPY = per member per year

Source: Adapted from Winkelman and Mehmud (2007), tables IV.1, IV.5, IV.7, and IV.8.

shelf" to predict similar outcomes in a population roughly comparable to that used to develop the model. Often a full-scale model recalibration is impractical. The relative similarity in R^2 values across the methods suggests that factors beyond statistical performance, such as transparency of the algorithms, clinical meaningfulness, feasibility, cost, availability, and appropriateness to purpose, should guide selection of the model used for prediction.

The second-to-last column of Exhibit 10.1 shows the R^2 values from models that make predictions for the same data used to derive model parameters (i.e., concurrent predictions). Concurrent models have much higher R^2 values than prospective models do because people with specified diagnoses in year 1 incur costs in that same year for observed diagnoses recorded during medical encounters. In contrast, whether someone who received a diagnosis in year 1 uses any health care in year 2, and how much care that person uses, is subject to chance. High R^2 values from concurrent models basically reflect the models' ability to predict (or explain) the past. R^2 values from concurrent and prospective models are not comparable.

Although R^2 is a valuable summary measure of model performance, it provides little intuitive feel for a model's ability to discriminate among cases with high and low values for the dependent variable. For such insight, we suggest examining actual and predicted outcomes within deciles of predicted outcomes. Exhibit 10.2 illustrates such outcomes among approximately 18,000 pneumonia patients (Iezzoni et al. 1996c); cases with a length of stay (LOS) over three standard deviations from the mean on a log scale were removed from the analysis. As seen in Exhibit 10.2, for APR-DRGs, which had an R^2 of 0.147, the predicted mean LOS of patients in the lowest decile was 5.5 days; the predicted mean LOS of patients in the highest decile was 12.8 days, slightly more than twice as long. For Refined DRGs (RDRGs), which had an R^2 of 0.170, the comparable LOS values were 4.1 days and 13.2 days, more than a threefold difference. Exhibit 10.2 also shows the actual mean LOS values in each decile. In general, predicted mean LOS values in the extreme deciles were closer together than were actual mean LOS values in these deciles. Deciles-of-risk tables such as Exhibit 10.2 usefully portray models' ability to discriminate among different types of cases. We revisit this topic in more detail later in the chapter when considering dichotomous outcomes.

Minimum and Maximum Values for R^2

In development (i.e., fitting) data, the R^2 of a linear model can never be less than zero. However, a model developed elsewhere may, when applied to new data, result in *SSE*s that are larger than the *SST*s, making R^2 negative. For example, suppose that average outcomes for the development and validation data set, designated as $\overline{Y}_{dev}$ and $\overline{Y}_{val}$, are different. A model that predicts $\overline{Y}_{dev}$ for every case in the validation data set will yield a negative R^2. More generally, a

negative R^2 will result whenever the $PRED_i$s determined on the development data are further from the Y_is (as measured by the sum of squared differences) than the constant prediction $\overline{Y}_{val}$ is from the Y_is. This result strongly indicates poor predictive power.

R^2 achieves its maximum value of 1.0 (100 percent) when all predictions are perfect—that is, when every $PRED_i$ equals the actual Y_i. However, for health outcomes, achieving R^2 values anywhere near 1.0 is unrealistic. For a given set of risk factors in a specific population, analysts can compute the maximum achievable R^2. For an illustration, consider a situation in which all

	Decile of Predicted LOS				
Severity Adjustment Method	**1**	**2**	**9**	**10**	R^2
	Mean Actual LOS (Mean Predicted LOS)				
AIM[a,b]	5.1 (5.5)	6.6 (6.8)	11.1 (10.9)	13.2 (12.8)	0.140
APR-DRGs[b]	5.0 (5.5)	6.6 (6.7)	11.6 (10.9)	13.0 (12.8)	0.147
Disease Staging Relative Resource Scale	5.1 (5.3)	6.4 (6.6)	11.2 (11.2)	13.4 (13.6)	0.144
PMC Resource Intensity Scale[c]	5.2 (5.2)	6.4 (6.8)	11.0 (11.2)	12.6 (12.9)	0.122
RDRGs	4.0 (4.1)	6.2 (6.5)	11.4 (11.5)	13.2 (13.2)	0.170
Disease Staging mortality probability	5.2 (5.3)	6.6 (6.9)	10.8 (11.0)	12.5 (12.6)	0.107
PMC severity score	5.1 (5.3)	6.5 (6.6)	10.8 (11.0)	12.6 (12.8)	0.122
Body Systems Count	5.1 (5.1)	6.5 (6.7)	11.0 (11.3)	13.0 (13.1)	0.133
Comorbidity Index	5.2 (5.3)	6.7 (6.9)	10.9 (11.0)	12.2 (12.5)	0.103
MedisGroups	5.2 (5.2)	6.5 (6.9)	10.7 (11.1)	12.4 (12.6)	0.109
Physiology score	5.1 (5.3)	6.7 (7.0)	10.6 (11.0)	12.1 (12.4)	0.099
Age, sex, and DRG only	5.2 (5.3)	6.8 (7.0)	11.0 (11.0)	11.9 (12.4)	0.097

EXHIBIT 10.2 Actual and Predicted Mean Lengths of Stay (LOSs) Within Deciles of Predicted LOS and R^2: Pneumonia Cases and Trimmed Data

[a]AIM = Acuity Index Method

[b]From a model that included interaction terms between APR-DRGs and stage

[c]PMC = Patient Management Category

Source: Adapted from Iezzoni et al. (1996c).

predictor variables are categorical. If a model includes every possible combination of predictor variables, the model is "fully saturated" (i.e., it includes all interactions). In this situation, each case falls uniquely into one of the cells defined by the interactions. For example, assume a model includes a sex and risk category, which can take one of five values. A fully saturated model would usually include ten (i.e., 2 × 5) variables, a constant term, plus nine dummy variables, which are summarized by saying there are nine degrees of freedom (df). One cell, for instance, would contain the data for all women in risk category 2.

In a fully saturated model, the predicted value of each case equals the average value of the dependent variable for all cases in the same cell. Linear regression models assume that the variance of the outcome variable is identical for different combinations of independent variables (i.e., risk factors). Under this assumption, cases in each cell have the same variance around their cell average. The highest achievable R^2 for any model that cannot make finer distinctions than those defined by these ten cells is expressed by the following equation:

$$max\ R^2 = 1 - \left(\frac{average\ variance\ of\ the\ outcome\ within\ each\ cell}{variance\ of\ the\ outcome\ in\ the\ whole\ population} \right).$$

This value can be estimated by determining

$$1 - \frac{(n-10)s^2}{SST},$$

where n is the number of cases in the data set and s^2 is the pooled estimate of the common variance of the outcome variable among cases in each cell (see Chapter 12 for details about these calculations, described in the context of provider profiling). No model can eliminate the variability measured by s^2 unless it uses additional information to further distinguish cases within each cell.

The more interaction terms a model includes, the closer R^2 will come to its maximum. Comparing R^2 to maximum R^2 can reveal, for example, how much explanatory power is sacrificed when risk is defined generically across conditions rather than condition by condition. For instance, consider APR-DRGs, which assign a separate stage for cases within each APR-DRG. Models must include separate independent variables for each stage within each APR-DRG because the stages for different conditions are not comparable. If APR-DRG and stage are the only variables included in a model, models should include interaction terms between APR-DRGs and stage. If the only data available for modeling are the APR-DRGs and their stages, R^2 and maximum R^2 are identical.

In contrast, consider a risk adjuster that assigns individuals to five risk categories independently of diagnosis. When using this risk adjuster and data

from ten diagnoses to predict an outcome, for example, analysts may wonder whether to (1) include a separate interaction term for each level of risk within each diagnosis or (2) include only nine separate variables to distinguish among the ten diagnoses plus four variables to distinguish the five risk categories. Modeling the first way uses 49 df, while the second uses only 13. Under the second strategy, R^2 is smaller than the maximum R^2 to the extent that the effect of risk category on the outcome varies by diagnosis.

The concept of maximum R^2 rarely appears in the published risk adjustment literature. Nonetheless, it is useful for two reasons. First, it helps distinguish poor fit caused by (1) use of a simpler model rather than one containing all possible interactions (measured by the difference between maximum R^2 and actual R^2) from that caused by (2) variation of the dependent variable within the cells defined by the independent variables used in modeling (measured by the difference between the maximum R^2 and 1.0) (Korn and Simon 1991). In a sense, the maximum R^2 measures how difficult an outcome is to model on a specific data set with a particular set of variables. Reporting maximum R^2 values for standard sets of variables might help characterize the difficulty of modeling on various data sets.

Grouped R^2

In the calculation of traditional R^2 (i.e., $1 - [SSE/SST]$), each difference between a person's actual and predicted costs contributes to the model's error and reduces the R^2. An individual's future health care costs are difficult to predict because much of health care spending results from events that are either totally unpredictable (e.g., being hit by lightning) or only somewhat predictable (e.g., persons with hypertension having a stroke). As a result, R^2 values for predicting health costs are lower than those that scientists studying more mechanistic outcomes usually see. For many purposes, however, we do not need accurate predictions for each person; we need only accurate predictions of the average costs for policy-relevant subgroups, such as persons enrolled in different health plans or receiving care from different delivery systems. Grouped R^2 measures models' ability to distinguish among **groups** of individuals with predictably different outcomes.

When building models to predict next year's Medicare costs, Ash and colleagues (1989) sought to demonstrate the importance of increasing R^2 by even a modest amount—from around 1 percent for a model using only demographic characteristics to about 6 percent for a model based on discharge diagnoses. Grouped R^2s helped show that the diagnosis-based models were especially useful for identifying **subgroups** of patients with predictably higher future costs.

Unlike the traditional R^2, a grouped R^2 is not an intrinsic property of a population. Rather, it depends on how the population is partitioned into

groups or "bins," such that each case belongs to exactly one group. Specifically, the grouped R^2 for a particular partition is defined as

$$Grouped\ R^2 = 1 - \left(\frac{GSS_{Error}}{GSS_{Total}}\right),$$

where $GSS_{Error} = \sum_{b \in B} n_b(\overline{Y}_b - average\ PRED\ in\ b)^2$, $\sum_{b \in B}$ is the sum over all bins, n_b is the number of observations in bin b, $\overline{Y}_b$ is the average value of Y in bin b, and

$$GSS_{Total} = \sum_{b \in B} n_b(\overline{Y}_b - \overline{Y})^2.$$

Looking at group totals (or averages) reduces random errors and highlights systematic differences between groups. The grouped R^2 imposes no penalty for predicting \$1,500 for each of two people when one costs \$0 and the other costs \$3,000, **if both belong to the same group.** Within the same group, positive and negative individual errors of predictions are allowed to cancel each other out.

Different binning rules yield different values of the grouped R^2 for the same model. Analysts should decide, preferably in advance, which binning approaches are of greatest interest. (This same issue applies to Hosmer-Lemeshow tests, described later in this chapter.) To see how well a model captures differences in next year's costs associated with differences in this year's costs, for example, we could form the bins as deciles of current costs. In addition to the usual dependence of performance measures on the data set used, grouped R^2 values depend heavily on the binning approach selected. If different risk adjusters are evaluated on the basis of bins formed using the adjusters' own predicted risk score, the bins will vary across risk adjusters. This difference is acceptable when risk adjusters' ability to predict outcomes within risk strata is informally evaluated. However, rigorous comparisons of risk adjusters should be based on a neutral binning method.

To determine how well models predict totals for policy-relevant subgroups, the 2007 SOA study examined predictive ratios (PRs) (Winkelman and Mehmud 2007). A model's PR for an outcome in a group is its predicted (or expected) average outcome for persons in that group divided by their actual (or observed) average outcome (Ash et al. 1989). In other words, PRs are expected-to-observed (*E/O*) ratios. SOA examined the PRs of models predicting total costs among persons within six specific conditions (asthma, breast cancer, diabetes, heart disease, HIV, and mental illness) and for groups defined by actual expenditures (next-year expenditures for evaluating prospective predictions and same-year expenditures for evaluating concurrent predictions). The condition groups considered all persons with the designated diagnosis (e.g., asthma), regardless of other medical problems. Because some people have multiple diagnoses and some have none, grouping people by condition does not assign each case into a single subgroup as is required

for computing a grouped R^2. Bins created by categories of total actual spending do put each case into only one group. However, if a bin has $0, or even near $0 costs, its *E/O* calculation is extremely unstable.

The SOA report contained a comment from Blue Cross Blue Shield of Mississippi stating that within the disease categories, average member predicted costs were very close to average member actual costs but that average male and female predictions within the disease categories were not as close to actual male and female costs (Winkelman and Mehmud 2007, 24). This finding is not surprising because interactions between sex and disease category were not explicitly included in any of the risk models. Any model makes better predictions for groups defined by variables it uses as predictors; when a model uses sex as a predictor, its average predictions are likely to be close to actual outcomes for all men and for all women. However, if separate variables for sex within each disease category were not included in the model, predictions and actual values will have no particular reason to be similar for, say, women with asthma.

For the same reasons, groups defined by a specific risk adjuster's classification system, by quantiles of that model's predicted risk, or by variables unique to that risk adjustment method yield grouped R^2 or PRs that are most favorable to that risk adjuster. Even a seemingly neutral approach, such as forming groups based on quantiles of the average *PRED* across all risk adjusters being compared, may be skewed. For example, imagine that models 1 and 2 produce nearly identical predictions and that model 3's predictions are equally powerful but somewhat different. If quantiles are based on the average of the three *PRED*s, the grouped R^2 will suggest that models 1 and 2 perform equally well and better than model 3 at capturing the differences in quantiles of average predicted risk. For this reason, in comparisons of one model's performance to another, grouped R^2s and PRs should be based on ways of grouping (or binning) that differ on characteristics intrinsic to the data, such as prior cost levels, rather than on groupings that are likely to advantage one model over another, such as deciles of a particular model's predicted risk.

Sometimes analysts use models to identify high-cost cases (also called "top groups") for disease management or other interventions. In this context, using R^2 to measure model performance does not directly address the relevant question: If each of several models is used to identify the same number of people in a top group, which one identifies the "best" group? Ash and colleagues (2001) illustrate some approaches to comparing models' ability to produce useful top groups—for example, those that contain (1) few incorrect selections (i.e., persons whose next year's costs would be low), (2) many good or accurate selections (i.e., persons whose next year's costs would be very high), and (3) many patients with potentially manageable disease (e.g., diabetes or asthma).

Handling Extreme Values of the Outcome Variable

Because the OLS algorithm chooses model parameters to minimize squared errors, extreme values of a continuous outcome can substantially affect parameter estimates. For example, health insurers may not care whether $10,000 in excess costs occurs because the observed cost of a single case exceeds the predicted cost by $10,000 or because the observed costs for ten patients each exceed expected values by $1,000. However, the squared term associated with one large error far exceeds the sum of squares associated with multiple smaller errors that add to the same amount. For example, $1 \times 10{,}000^2$ is ten times larger than $10 \times 1{,}000^2$. Health care costs often have distributions with "heavy" or long right tails due to a few extreme values. The presence of extreme values in a model-fitting data set can distort predictions for the "main body" of the data.

Health services researchers have several strategies for handling extreme (outlier) cases. One simple strategy, called *trimming,* is to drop cases with extreme values. The impact on R^2 of trimming outlier cases is not obvious. It effectively reduces both *SSE* and *SST*. Because R^2 reflects the ratio of these two values, the effect of eliminating outliers depends on the relative magnitude of the changes in the two terms.

Another strategy, called *top-coding, winsorizing,* or *truncation,* is to reset extreme observations to a less extreme value. (The nomenclature used to distinguish these approaches in the health services research literature is inconsistent.) In top-coding, the actual values for all cases with values greater than some number M are replaced by M. Thus, top-coding retains all cases but reduces the influence of outliers by drawing them toward the mean. Top-coding is especially suited to situations in which cases with particularly high values might be treated differently, such as when a reinsurer pays for any individual's costs that exceed a specified threshold. In winsorizing, equal numbers of both high- and low-value cases are replaced by specified values. For example, the five lowest values are reset to their maximum value and the five highest to their minimum. Winsorizing aims to create more stable estimates of means for essentially symmetrical distributions (Dixon and Tukey 1968; Dixon and Massey 1969).

Various studies involving risk adjustment have used top-coding. In deriving APACHE III models to predict ICU LOS, Knaus and colleagues (1993) capped all ICU lengths of stay at 40 days. As noted earlier, the SOA study of risk adjustment models analyzed the data three ways: by including all costs at their original values, by top-coding costs at $250,000, and by top-coding costs at $100,000 (Winkelman and Mehmud 2007). As expected, R^2s were lowest when costs were predicted using their original values and highest when costs were top-coded with the lower threshold (Exhibit 10.1).

While extremely expensive cases are rare, they are an important feature of health care spending. Thus, we prefer top-coding to trimming. Top-coding

typically eliminates a great deal of variability in the outcome measure while lowering its mean value by only a very small fraction. In descriptions of top-coded analyses, it is advisable to report both the fraction of the cases that exceeded the threshold and the fraction of total dollars lost due to top-coding. One simple way to put these lost dollars back into the analysis is to multiply all reported dollars by an inflator:

$$INFLATOR = \frac{mean\ of\ the\ original\ dollars}{mean\ of\ the\ top\text{-}coded\ dollars}.$$

Models built on top-coded data can make predictions that exceed the top-coding threshold, although they do so infrequently. While unsatisfying, such predictions are not a statistical problem. However, especially if the goal is to predict costs that lie below a reinsurance threshold, it makes sense to compare top-coded actual dollars to top-coded predictions that use the same threshold.

Another common strategy for limiting the influence of extreme-valued cases is to transform the dependent variable, for example, by replacing each outcome with its natural logarithm or some other member of the Box-Cox family of transformations (Box, Hunter, and Hunter 1978; Spitzer 1982; Atkinson 1985). Such transformations enable models to be built on an outcome whose distribution more nearly conforms to the normality assumptions that underlie OLS modeling. Transformation of continuous outcome variables is most appropriate when the goal is to identify statistically significant predictors of the outcome (to better meet the assumptions underlying statistical tests). However, such transformations are problematic when the goal is to accurately predict values of the original dependent variable, as discussed in the following paragraphs.

At a minimum, predictions based on a transformed outcome variable must be retransformed to the original scale before either calculating statistical performance measures or predicting outcomes. As noted earlier, in cost predictions, payments are made in dollars, not in "log dollars"; R^2 values from predicting the logarithm of an outcome and from predicting the actual outcome are not comparable. OLS maximizes R^2 for the data to which the model is fit, whether on an original or transformed scale. Regardless of the scale used to develop the model, the model's performance should be evaluated in the actual units of interest (e.g., dollars, days).

Duan's smearing estimator is a widely used, theoretically attractive approach to making predictions in the original scale (e.g., dollars rather than log dollars) using a model that was developed on log-transformed data (Duan 1983). In this approach, each log-dollar prediction—let's call it *PRED*—is first retransformed to the dollar scale by undoing the transformation—that is, by exponentiating it to get e^{PRED}. The problem is that the e^{PRED} values

are (collectively) too small (i.e., their sum is predictably less than total spending in the population). Thus, Duan's method uses predicted values of the form $k_s \times e^{PRED}$, where k_s is an averaged (i.e., "smeared") correction factor. To calculate k_s one finds the log-scale residuals ($log\ Y - PRED$), exponentiates each, and finds their average. With right-skewed data, it should be easy to verify directly that multiplying by k_s yields better estimates of the original Y values than using the e^{PRED}s without the correction factor does. This method is exactly right when the Ys are log-normally distributed and the relationship between the predictors and dollars is multiplicative. For example, if the presence of a risk factor typically adds 10 percent to costs irrespective of how expensive a person's care might be otherwise, adding 0.10 in the log scale and using the smearing estimator to make predictions in the original scale (dollars) could produce better predictions than would modeling dollars directly and adding a fixed amount (e.g., $500) to everyone's predictions. One nice feature of this approach is that it is easy to verify how well it worked. Because the purpose of the smearing estimator is to eliminate the downward bias in the e^{PRED}s, if the average of the $k_s \times e^{PRED}$ values is not approximately equal to the average of the Ys, it might be worth simply multiplying the e^{PRED}s by whatever constant is needed to get the averages to match.

Despite the theoretical appeal of modeling with transformed data, retransformed predictions often produce markedly lower R^2 values than do predictions from a model fit directly to the untransformed data, partly because the OLS algorithm specifically estimates parameters that maximize R^2 in the scale of interest and also because some errors matter much more than other errors that have the same size in the log scale. For example, although both of the following errors are comparable in the log scale, predicting $10 when $100 is spent (a factor-of-10 error that mispredicts by $90) matters much less in the original dollar scale than predicting $10,000 when actual spending is $100,000. In our study of hospital-based severity measures, we evaluated R^2 both on untransformed LOS data and log-transformed data, using Duan's smearing estimator to retransform predictions before calculating R^2. For both pneumonia cases and hip fracture cases, when LOS was used as the dependent variable, R^2 values were as high as or higher than they were when results were retransformed from a model using log (LOS) as the dependent variable (Iezzoni et al. 1996c; Shwartz et al. 1996). Lumley and colleagues (2002) used simulation to illustrate that OLS may be appropriate for modeling cost data, even for sample sizes as small as 500 cases.

As an alternative to transforming the data, the generalized linear model (GLM) provides a comprehensive framework for developing multivariable models with non-normally distributed data. GLM assumes that some function of the outcome, called a *link function,* can be modeled as a linear function of the predictors. GLM seeks parameters that predict outcomes in their natural scale directly, rather than in a transformed scale. The GLM framework also

allows for nonconstant variance, for example, as a function of the mean value of the outcome (because costs for cases expected to be expensive typically vary more than costs for cases expected to be inexpensive). Generalized linear models are described in a classic but sophisticated text by McCullagh and Nelder (1989), which assumes knowledge of likelihood functions. Hoffman (2004) provides simpler and more applied information about GLM. Barber and Thompson (2004) wrote a relatively nontechnical article illustrating the use of GLM to analyze cost data.

Another way to limit the influence of outliers is to seek models that minimize mean absolute prediction error, also called the mean absolute deviation (*MAD*), rather than the squared error. *MAD* is the average *absolute value* of *Y – PRED*. The *D* for "deviation" in *MAD* denotes the same quantity as "error" in the term *SSE*. Although standard software packages do not choose model coefficients that minimize *MAD*, analysts still can compare risk adjusters on the basis of *MAD*. SOA took this approach in evaluating risk adjustment models (Winkelman and Mehmud 2007), expressing *MAD* as a percentage of average PMPY (per member per year) cost. The last column of Exhibit 10.1 shows *MAD* values from the SOA report. The actual report uses the term *MAPE* (mean absolute prediction error) rather than *MAD*. For this exhibit, models are used "out of the box"—that is, the vendors supplied the coefficients (the results from refitting the models to the SOA data were not used) to predict next year's annual spending, with claims top-coded at $250,000. Usually, better models have higher R^2s and lower *MAD*s. However, relative model performance occasionally differs by statistical performance measure. For example, DxCGs have the highest R^2, but RxGroups and PRGs have lower *MAD* values; also, RxGroups and PRGs have only slightly lower R^2 values than DxCGs have but noticeably lower *MAD*s.[11]

MAD may seem a more natural measure of model fit than R^2 is. However, a model has the highest R^2 when the sum of its squared errors (*SSE*) is smallest. When predicting averages for moderately large groups, models that minimize *SSE* predict more accurately than do models that minimize *MAD*.[12]

Finally, the transparency of variable transformations and nonlinear models is important. Developers must convince nontechnical audiences that their methods make sense (have face validity; see Chapter 9). They also need to engage others who are particularly knowledgeable about the specific context of the work in critiquing the model. OLS modeling is fairly easy to explain, and as noted earlier, Lumley et al. (2002) have shown that it generally performs well.

Interpreting R^2

With experience, analysts learn what values of statistical performance can be expected for predictive models and what values are "good enough" for models to be successful. For example, more than two decades ago, actuaries were

satisfied with fairly simple demographic models with R^2 values less than 0.05 when using data from the current year to predict the following year's health care costs for a Medicare enrollee (Ash et al. 1989; Epstein and Cumella 1988). Small R^2s indicate that individual *PRED*s typically differ from observed outcomes by almost as much as they would if each *PRED* always equaled the population mean.[13] Even so, such *PRED*s can identify important systematic differences in moderately sized groups and can facilitate finding individuals with expected high costs who might be good candidates for case management.

The SOA study (Winkelman and Mehmud 2007) demonstrated that validated R^2s are about 20 percent for models predicting next year's actual costs from this year's diagnoses or pharmacy claims (plus age and sex). However, with more clinical information—for example, laboratory values, functional status, and health behavior data, which are typically not available in administrative databases (see Chapter 5)—prospective models might produce larger R^2 values. As discussed in the next section, addition of laboratory values to models that predict death has significantly increased model performance, suggesting a potential for improved cost prediction. Many observers are unimpressed by models that explain 15 to 20 percent of the variation in cost. However, context is all-important. Models predicting next year's actual costs for individuals may never explain most of the variability because current data cannot identify which individuals will incur catastrophic expenses next year. Nonetheless, prospective models in use today can distinguish **subgroups** of persons whose predicted costs next year differ by factors as high as 100 to 1.

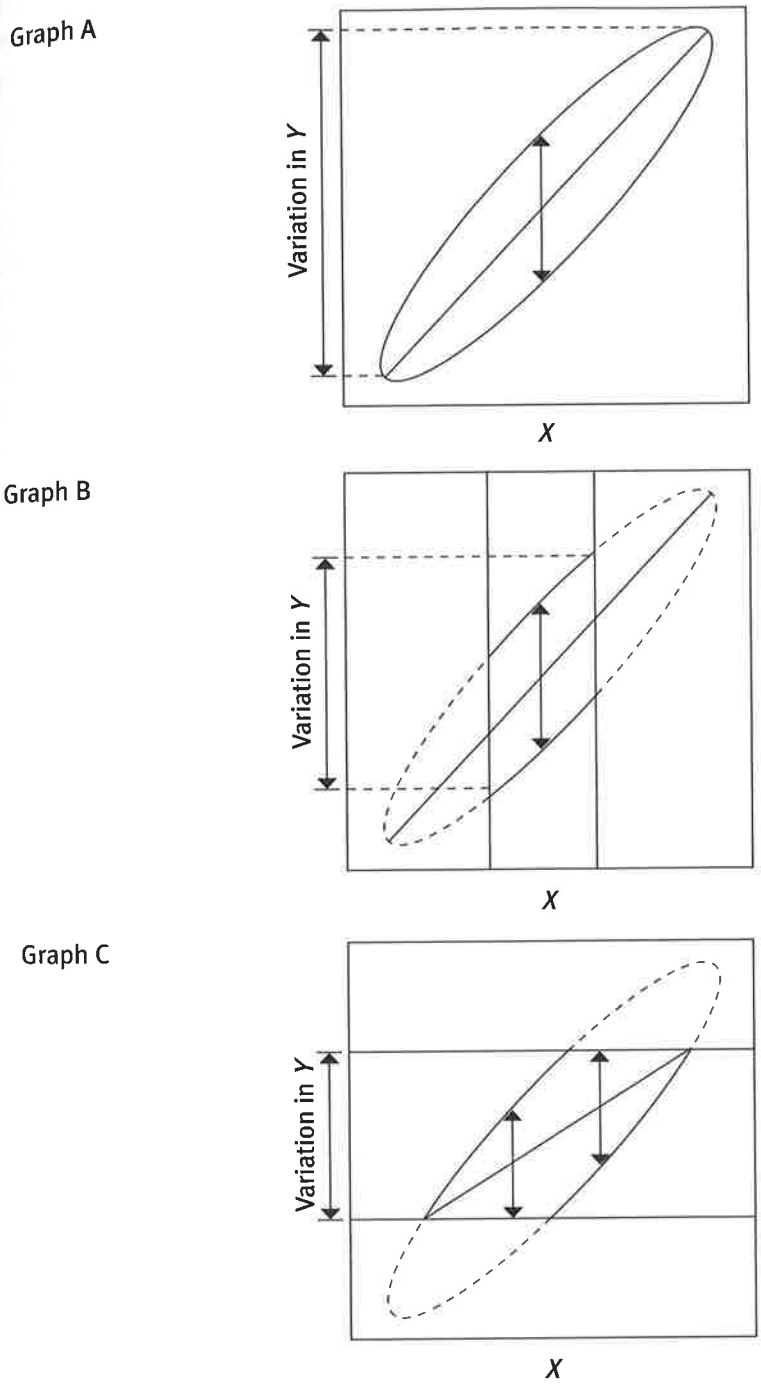
R^2 values depend on the cases in the database used and on the risk factors available for modeling. The dispersion of the independent and dependent variables significantly affects statistical performance, particularly R^2. Exhibit 10.3 shows three schematic diagrams of models that predict Y from a single variable X, where larger values for X indicate greater risks for poor outcomes. Graph A shows the classic bivariable normal situation. R^2 is approximately equal to

$$1 - \frac{variance(Y - PRED)}{variance(Y)}.$$

Graphs B and C in Exhibit 10.3 show what happens when only some of the data are available for modeling. In graph B, cases with extreme values of X are eliminated, which produces less variation in Y but no change in the variation in $Y - PRED$ for the remaining cases. In this situation, R^2 should decrease. In graph C, cases with extreme values of Y are removed, which reduces both kinds of variation. The net effect on R^2 is unpredictable.

Another example illustrates how the characteristics of the data set affect statistical performance. Suppose two risk adjusters are applied to two different data sets and are equally accurate in predicting outcomes for cases across the range of independent variables (i.e., SSE/n is identical for each risk adjuster).

EXHIBIT 10.3 How Differences in the Modeled Data Affect R^2



Note: Graph A shows a regression line fit to a schematic scatterplot of bivariate normal data. In graph B, cases with extreme values of X have been removed; in graph C, cases with extreme values of Y have been removed. The shorter, vertical arrows in the body of each graph indicate variation in $(Y - PRED)$. The smaller the ratio of variation in $(Y - PRED)$ to variation in Y, the larger R^2 is.

The method applied to the data set with more variability in the outcome variable generates a higher *SST* and, thus, a higher R^2 (Korn and Simon 1991). The higher R^2 yields misleading conclusions about which risk adjuster performs better.

Certain data sets are formed by oversampling particular types of cases (e.g., cost outliers, patients who died) to increase the amount of information available about them. Average outcomes for these oversampled cases often differ greatly from those for the general population. As a result, the variability of the dependent variable is artificially increased, which usually leads to an R^2 higher than that for the general population. When sampling weights are known, they can be used to adjust for this difference; in this case, the R^2 from a weighted least squares regression suggests how the model will perform when applied to a general population.

Because of the sensitivity of R^2 to characteristics of the data set, some conclude that R^2 is unsuitable for comparing models developed on different data sets (Cox and Wermuth 1992). More generally, comparisons of models developed on different data sets can be misleading regardless of the measure of model performance used.

Importance of the Form of the Risk Score for Interpreting R^2

Some risk adjustment approaches rate cases by assigning categories of risk, while others yield scores with continuous values. To capture complex nonlinear relationships between risk and outcome, analysts sometimes recode continuous scores as discrete categories. The approach used to create categories can also affect R^2.

One approach involves distributing cases relatively evenly across groups (e.g., by defining five quintiles of increasing risk). However, in a population in which much of the complexity, cost, and potential for cost savings is concentrated in a small fraction of cases, it may make more sense to use unequal bucket sizes, letting the top one or two categories contain fewer cases than the lower-risk groups: perhaps only 5 percent of the cases, or when sample sizes allow, 1 percent or less. Especially when costs are not top-coded, models that predict from categories that use fine gradations at the high end of risk perform better than those that divide the population into a similar number of equal-sized buckets.

The relative merit of different approaches to defining categories may depend on the context and purpose. For example, suppose two risk adjustment methods each divide a population with a 4.5 percent death rate into four risk categories. Across method A's four categories, 1, 3, 4, and 10 percent of the cases die, while across method B's four categories, 0.5, 4, 5, and 25 percent of the cases die. At first glance, method B seems to discriminate better. Suppose, however, that 25 percent of the cases fell into each of method A's four categories, while only 1 percent of the cases were in each

of method B's lowest- and highest-risk groups and 49 percent in each of the two middle risk categories. For many purposes, method A is more useful because it can find fully one-fourth of the people with ten times the risk (i.e., the 10 percent death rate category) of another one-fourth of the population (i.e., the 1 percent death rate category). In contrast, method B places fully 98 percent of the population in categories with very close to average risk. However, if the goal is to find the 1 percent of the population at highest risk, method B is more useful. More generally, with categorical risk scores, better models assign larger fractions of the population to risk categories with either very low or very high average outcomes.[14]

Measures of Model Performance for Dichotomous Outcomes

To demonstrate the measurement of model performance for dichotomous outcomes, we compute a score measuring risk of in-hospital death and set a cutoff score such that we predict patients will die if their scores exceed this cutoff. We predict that patients with lower scores will live. We array such data in a two-by-two table as shown in Exhibit 10.4, thereby defining several useful quantities:

- True positive cases = A
- False positive cases = B
- True negative cases = D
- False negative cases = C
- Prevalence = $(A + C)/(A + B + C + D)$
- Sensitivity = $A/(A + C)$
- Specificity = $D/(B + D)$
- Predictive value positive = $PV^+ = A/(A + B)$
- Predictive value negative = $PV^- = D/(C + D)$

From different perspectives, the last four quantities measure the percentage of cases correctly classified. *Sensitivity* is the percentage of deaths

	Patient Outcomes		
Risk Score Prediction	**Dead**	**Alive**	**All Cases**
Dead	A	B	A + B
Alive	C	D	C + D
All cases	A + C	B + D	A + B + C + D

EXHIBIT 10.4 Layout for Comparing a Dichotomous Outcome to a Dichotomized Prediction

correctly classified by the prediction rule, while *specificity* is the percentage of live patients correctly classified. The *predictive value positive* is the percentage of patients predicted to die who are classified correctly, while the *predictive value negative* measures the percentage of patients predicted to live who are classified correctly.

The cases used in the denominators of PV^+ and PV^- depend on the classification rule, which complicates comparisons of different classification rules based on PV^+ and PV^-. In contrast, sensitivity (the *true positive rate*) and specificity (the *true negative rate*) use denominators defined by the death rate within the study population, which makes sensitivity and specificity more suitable for comparing different classification rules involving the same population. The *false negative rate* is defined as 1 – *sensitivity*, and the *false positive rate* is 1 – *specificity*.

Parker and colleagues (2006) compared the prevalence of various risk factors determined from clinical data on 38,230 patients receiving CABG surgery in California in 2000–2001 to risk factors determined from administrative data. Describing the level of success associated with identifying true cases of cardiogenic shock (as identified in clinical data) from diagnoses coded on claims, they reported:

- Sensitivity = 0.484
- Specificity = 0.998
- Predictive value positive = 0.82
- Predictive value negative = 0.98

Thus, although only 2 out of 1,000 cases (i.e., [1 – 0.998] × 1,000) were misclassified from the administrative data as having cardiogenic shock when in fact they did not, only about 48 percent of patients identified from clinical data as having cardiogenic shock could be identified from administrative data. Also, while fully 98 percent of patients not coded with cardiogenic shock did not have it, among people with administrative codes indicating shock, this condition could be confirmed in the clinical records only 82 percent of the time.

Evaluations of diagnostic rules in clinical practice often rely on two-by-two classification tables. In this context, the prevalence of a condition in a population strongly affects the predictive values of a test. For example, the PV^+ and PV^- values for cardiogenic shock were from a population in which the prevalence of cardiogenic shock was about 2 percent. With a condition this rare, prediction of nonoccurrence is likely to be correct while accurate identification of the cases in which it occurred is more difficult. An identification method applying the same sensitivity and specificity to a population in which the outcome occurred as often as 10 percent of the time would cause PV^+ to increase to 0.96 and PV^- to decrease to 0.5.

Sensitivity and specificity thus describe properties of tests in the abstract, whereas PV^+ and PV^- address the consequences of using a particular test in a specified population. A major problem with using sensitivity and specificity determined from two-by-two tables to measure model performance is the need to select a single cutoff to dichotomize the risk score. For some purposes, defining the cutoff could closely mimic the measurement methodology to be applied, such as the use of scores to identify a manageable number of high-risk patients to receive special follow-up. However, even in those cases, the best cutoff is situation-specific and depends on such factors as the funds available to monitor the selected patients.

Typically, no single cutoff is unequivocally best, which makes two-by-two tables poorly suited for comparing the performance of risk adjustment methods. When different researchers describe the performance of risk adjusters, each using a different cut point, reported sensitivities can vary widely, impeding comparison of specific values. When cut points are chosen to produce identical sensitivities, one can compare specificities. However, judgment of which model is better may depend on the value of the sensitivity used to establish the cut point. For example, one model may accurately identify a few cases for whom death is very likely but not discriminate well among the general population; another model may identify a substantial fraction of cases whose risk of death is much higher than average but still unlikely. In language defined later in this chapter, such inconsistent judgments can arise whenever the receiver operating characteristic (ROC) curves for different risk adjustment models cross.

Models for Predicting Dichotomous Outcomes

Logistic regression is typically used to model dichotomous dependent variables—that is, those that take the value 1 when an event occurs and 0 when it does not. In logistic regression, the dependent variable is the natural logarithm of the odds of the event (written as ln O and called "log odds" or "logit"). If p is the probability that an event occurs and $q = 1 - p$, the odds of the event is p/q. For example, if the chance of an event is 3 in 4, $p = 3/4$, $q = 1/4$, and the odds of the event is 3 to 1 or 3.0. If p_i is the probability of the event for the i^{th} case, the logistic regression model has the form

$$\ln(O_i) = \ln(p_i/(1 - p_i) = a + \Sigma_j b_j X_{ij}.$$

As in multiple regression models for continuous outcomes, the X_{ij}s may be dummy variables representing risk classes, covariates plus the risk classes, covariates plus a single risk score, or a subset of significant variables identified through a selection procedure, such as stepwise regression. With a continuous outcome variable, such as cost, the predicted value for the i^{th} person, $PRED_i$, is an estimate of that person's expected cost. In logistic regression, the predicted value also estimates the expected value of the outcome for a

person—that is, in this case, the outcome is a 0/1 variable and we denote $PRED_i$ by p_i, the probability that the event occurs (outcome = 1) for the i^{th} person. Note that single observations can be only 0 or 1; they never equal their expected value p_i, which is always a number larger than 0 and less than 1.

From an estimate of ln O_i (written ln $\hat{O}_i$), the predicted probability of the event for the i^{th} case is

$$PRED_i = \frac{e^{\ln O_i}}{1 + e^{\ln O_i}} = \frac{1}{1 + e^{-\ln O_i}}.$$

Methods other than logistic regression may be used to estimate the numbers $PRED_i$, such as binary splits algorithms, probit models, ordinary least squares models, Cox proportional hazards regression models, or determinations by panels of clinical experts. Many measures of model performance, including the traditional R^2 and the c-statistic (see discussion later in this chapter), may be calculated from the set of pairs of numbers, one for each case, consisting of (1) the estimated probability of the event of interest for that case—that is, its $PRED_i$, and (2) the actual outcome Y_i, coded as either a 0 or a 1.

Another measure of model fit—the likelihood (L) of the observed data as predicted by the model—is also a function of these pairs. If the outcome of interest is death, L for a particular model is computed by multiplying together the predicted probabilities of living 1 – *PRED* for each case who lived and the predicted probabilities of dying *PRED* for each case who died. Better models have larger 1 – *PRED* values for cases who lived and larger *PRED*s for those who died. Thus, they have larger likelihoods. The Akaike Information Criterion (AIC) uses the likelihood to measure the deviance between the actual data and the model, building in a penalty for model complexity (Akaike 1973). Neither the likelihood nor the AIC is standardized to a common, interpretable scale, such as the 0 to 1.0 interval containing the c-statistic (see discussion later in this chapter). One standardized likelihood-based measure is a generalization of R^2. Similar to the traditional R^2, it summarizes the improvement in model fit for the full model over a model with just an intercept (Nagelkerke 1991). Likelihood-based measures of model performance are sometimes reported in the health services research literature but less often than the traditional R^2 and c-statistics are reported.

As is true of OLS modeling for predicting continuous outcomes, logistic modeling is based on certain assumptions, most important of which is that ln(*Odds*) can be expressed as a linear function of the predictors. Results from logistic models are most clearly valid when these assumptions are approximately true. Nevertheless, both logistic and OLS modeling are performed when these assumptions are clearly not met because both techniques are generally robust to fairly large departures from the assumptions. *Robust* means that key findings are not much influenced by the mismatch between the ideal and the real data to which the models are applied. More important, risk adjustment models are

typically useful to the extent that their predictions match reality rather than the extent to which the models meet underlying assumptions. When employed carefully, these modeling procedures yield reasonable predictions.

Model Calibration and Discrimination

As discussed earlier, models are said to be calibrated to a data set when the average of their predicted values matches the average of the actual outcomes. Overall calibration error is measured as $AVE(Y) - AVE(PRED)$, where $PRED$ is the predicted value of Y. For a dichotomous outcome, Y_i records the presence of the outcome event: 1 if the outcome occurs and 0 if it does not. With this coding, the average Y is the actual outcome rate in the population, and the average $PRED$ is the average predicted outcome rate for the same group. In OLS regression, the calibration error of its predictions is zero in the data set to which the model was fit. If some other fitting approach is used (e.g., the maximum likelihood method employed by logistic regression software), average actual and predicted outcome rates may differ, although usually by small amounts.

For a given outcome such as death, model discrimination represents the extent to which a model predicts higher probabilities of death (higher *PRED*s) for patients who died than for those who lived. Discrimination can be portrayed graphically by drawing two histograms of the predicted probabilities on the same scale: one histogram for patients who die and another for those who live. The histogram of *PRED*s for patients who lived should lie to the left of the histogram for patients who died. Exhibit 10.5 illustrates two such histograms, scaled for a situation in which four times as many cases lived

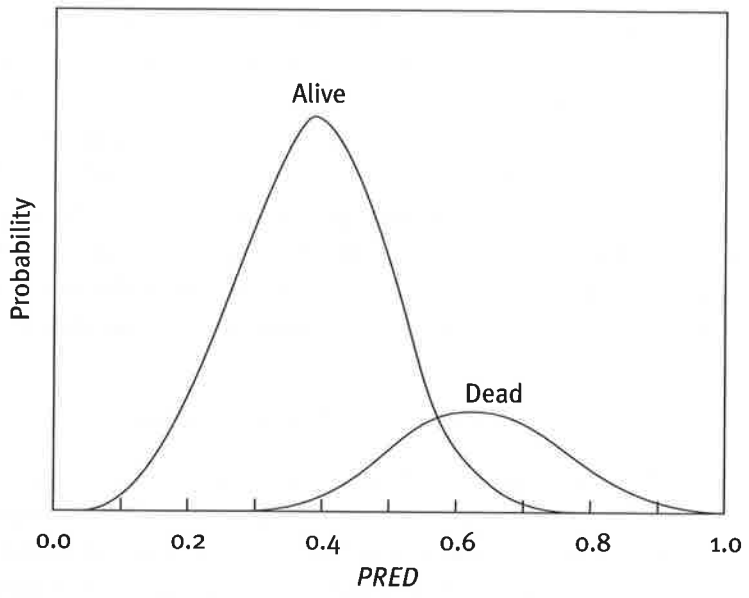


EXHIBIT 10.5 Schematic Drawing of a Covariance Graph: Comparative Histograms for *PRED* by Actual Outcome

as died (i.e., a 20 percent death rate). This type of drawing is called a covariance graph (Yates 1982). With better discrimination, the two histograms overlap less.

While good model calibration is desirable, discrimination is more fundamental. To see why, imagine a population with a 10 percent death rate. A model that assigns $PRED = 0.1$ to every person is perfectly calibrated, but it cannot distinguish patients who live from those who die. On the other hand, suppose that a second model assigns $PRED = 0.8$ to every patient who lives and $PRED = 0.9$ to the patients who die. Although the second model produces entirely wrong numerical predictions, the model can perfectly predict the outcome. While recalibration can fix the problem with the second model, no simple change can make the first model useful for identifying patients who die.

The c-Statistic as a Measure of Model Discrimination

Although some disagreement remains, most analysts believe that the c-statistic is valuable for measuring performance of models predicting dichotomous outcomes (Harrell et al. 1984). The c-statistic has several equivalent definitions. For an illustration of one definition, consider all possible pairs that can be formed, such that one patient dies and the other lives. The c-statistic equals the proportion of pairs in which the predicted probability of death is higher for the patient who died than for the patient who lived. Each tied pair counts as one-half.

The c-statistic also equals the area under the receiver operating characteristic (ROC) curve. ROC curves result from converting the information in a covariance graph into multiple two-by-two tables, such as Exhibit 10.4, and then plotting *sensitivity* versus 1 – *specificity* from each table. The following example illustrates the ROC curve using covariance graphs. First, the two histograms are converted to distributions by rescaling each curve so that it has an area of 1.00. Given a predicted probability of death for each case ($PRED_i$) and a specified cutoff (t), we make a rule as follows: If $PRED_i$ is greater than t, patients are expected to die; if $PRED_i$ is less than t, patients are expected to live. Exhibit 10.6 displays the consequences of this rule for a particular value of t. The true positive rate (i.e., sensitivity) is the area to the right of t and under the curve for the patients who die. The false positive rate (1 – *specificity*) is the area to the right of t and under the curve for the patients who live. In this way, each t is associated with a point determined by a pair of numbers:

$$P(t) = (x(t), y(t)) = (1 - specificity(t), sensitivity(t)).$$

As marked in Exhibit 10.6, the cut point t identifies a particular point $P(t)$ on this curve. The ROC curve is the set of points $P(t)$ that are traced on a unit square as t varies from its lowest to its highest value (Exhibit 10.7). When t is zero, all cases are declared positive and the pair (1,1) is generated.

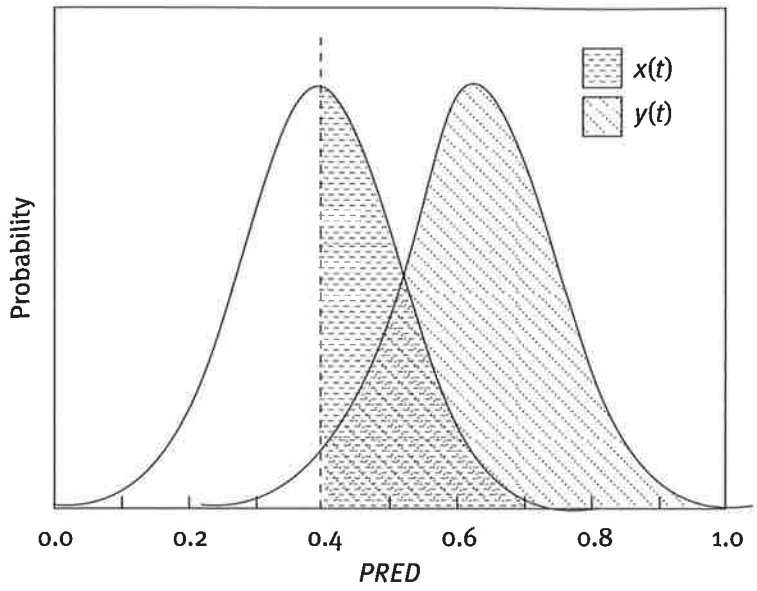


EXHIBIT 10.6
Distributions of *PRED* by Actual Outcome with Indications of True and False Positive Rates as Determined by a Cut Point *t*

$x(t)$ = fraction of cases that live for which $PRED > t = 1 - specificity$
$y(t)$ = fraction of cases that die for which $PRED > t = sensitivity$
Note: The left-hand curve = alive; the right-hand curve = dead.

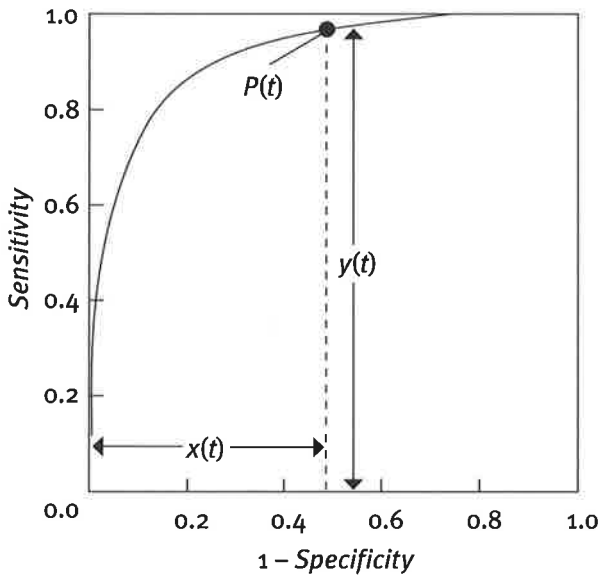


EXHIBIT 10.7
ROC Curve Associated with Distributions Shown in Exhibit 10.6

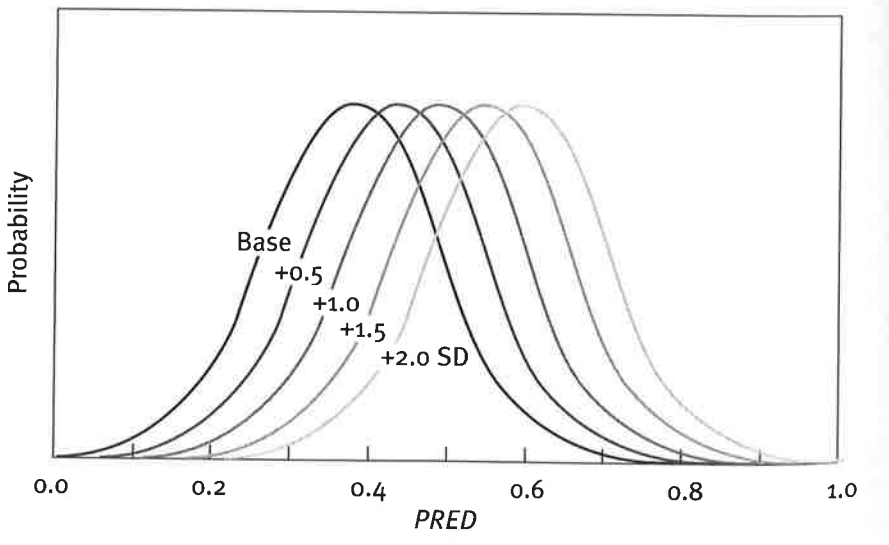
Note: $P(t) = (x(t), y(t))$ is determined using the particular cut point t shown on Exhibit 10.6. As t increases from 0 to 1, the ROC curve is traced out, starting in the northeast corner (1,1) and finishing in the southwest (0,0).

As the cutoff t increases, the vertical line in Exhibit 10.6 shifts to the right, and fewer cases are called positive. Thus, both the true and false positive rates decrease, and the new point lies to the left of and lower than the previous point. When t is 1.0, all cases are declared negative and the pair (0,0) is generated.

A model with good discrimination has a high true positive rate and low false positive rate, which generates a curve that passes close to the upper-left corner of the plot—the point (0,1) in Exhibit 10.8. The area under this curve, which is equivalent to the c-statistic associated with the model generating the covariance graph, is close to 1. Most risk adjustment models produce fairly continuous distributions of predicted values. When they do so, drawing a ROC curve is relatively straightforward. However, models with a small number of categorical independent variables generate only a few pairs of $(x(t), y(t))$. In these circumstances, alternative approaches can estimate the presumably continuous underlying ROC curve (Centor and Schwartz 1985).

The relationship between c and the rank sum statistic, s, offers more insight into what the c-statistic measures. s tests the following null hypothesis: The median value of *PRED* is identical for patients who live and those who die (which implies the model has no ability to discriminate). s is formed by combining the n_0 cases who live with the n_1 cases who die and assigning ranks according to the values of *PRED*. The case with the lowest value of *PRED* is assigned rank 1, the case with the second lowest value is assigned rank 2, and so on, up to $n_0 + n_1$, the rank assigned to the case with the highest value of *PRED*. For cases with the same value of *PRED*, the ranks are averaged. For example, if four cases share the same lowest value of *PRED*, each is assigned 2.5—that is, (1 + 2 + 3 + 4)/4.

EXHIBIT 10.8 Overlapping Normal Distribution Curves Displaying Various Amounts of Shift



The rank sum statistic s results from adding the ranks of all n_1 patients who die. The smallest possible value for s is achieved when all patients who die have a predicted probability of death (*PRED*) that is less than *PRED*s for cases who live. Therefore, $s_{min} = 1 + 2 + \ldots + n_1$. Alternatively, $s_{min} = n_1(n_1 + 1)/2$. This model discriminates perfectly, but it is wrong; the persons who die have the lowest predicted probabilities of death and those who live have the highest. The largest possible value of s is achieved when the n_0 cases who live have ranks 1 to n_0, leading to $s_{max} = (n_0 + 1) + (n_0 + 2) + \ldots + (n_0 + n_1)$. The sum of these n_1 ranks is $s_{min} + n_0 n_1$. Such a model also discriminates perfectly, and it is right; the persons who die have the highest predicted probabilities of death. Whatever the value of s, c is the following unique linear function of s:

$$c = \frac{(s - s_{min})}{(s_{max} - s_{min})}.$$

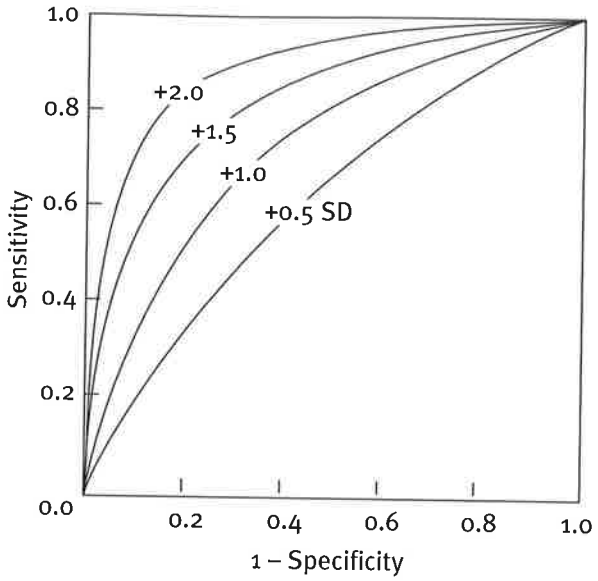
When $s = s_{min}$, $c = 0$; when $s = s_{max}$, $c = 1$. When the average rank of *PRED* is identical for cases who live and those who die (i.e., when s equals s_{mid}, a number halfway between its minimum and maximum), $c = 0.5$. The c-statistic depends only on the ranks of predictions, not on their actual values. Thus, the value of c is unaffected by how close individual $PRED_i$s are to their Y_is or by how well the average prediction, *AVE*(*PRED*), matches with *AVE*(*Y*); as a result, it provides no information about model calibration.

In summary, the c-statistic measures model discrimination, achieving its maximum value of 1.0 when all *PRED*s for cases with the outcome are larger than any *PRED*s for cases without the outcome. When models have no ability to discriminate (e.g., probabilities are randomly assigned to cases with and without the outcome), the expected value of the c-statistic is 0.5. A c-statistic much less than 0.5 indicates model discrimination but improper coding of the risk scores. This situation is unlikely to happen, but if it did, it could easily be remedied by reassigning risk scores (values of *PRED*) that go in the other direction (e.g., each person's *PRED* could be rescored as 1 – *PRED*).

To suggest how various c values reflect different levels of discrimination, Exhibit 10.8 shows overlapping normal distribution curves when the distribution for the patients who die is shifted to the right of the distribution for patients who live by 0.5, 1.0, 1.5, and 2.0 standard deviations. Exhibit 10.9 shows the four ROC curves for the shifted distributions. The values of c associated with these shifts are 0.64, 0.76, 0.86, and 0.92, respectively.

The c-statistic and ROC curve derive from measures of sensitivity and specificity. As noted earlier, these measures do not depend on the prevalence of the condition (e.g., the death rate in the population), a fact that has raised questions about c-statistics. Hilden (1991) argues that the value of a diagnostic test to clinicians depends on the prevalence of the targeted outcome in their patient population; hence, the insensitivity of c-statistics and ordinary ROC curves to prevalence is a weakness. He proposed using ROC-like

EXHIBIT 10.9
ROC Curves for Four Different Amounts of Shift Between *PRED*s for Cases Who Live and *PRED*s for Cases Who Die



Note: The areas under the four curves are 0.64, 0.76, 0.86, and 0.92 for shifts of 0.5, 1.0, 1.5, and 2.0 SDs, respectively.

pictures scaled for prevalence (thus affecting total numbers of misclassified cases) and possibly for different "costs" of false positive versus false negative errors to the patient or society. Such graphs highlight the trade-offs associated with using different cut points to decide where to call a case positive. Hilden (1991) also provides a contrived but nonetheless disturbing example of two procedures that are logically equivalent yet lead to different ROC curves and different c-statistics.

Knaus and colleagues (1991) reported a c-statistic of 0.90 for predicting death using APACHE III, which was based on the same 17,440 ICU patients from 40 hospitals used to develop the model. They also reported sensitivity, specificity, predictive value positive, and predictive value negative for three different cut points used to predict whether a patient would die: model-predicted probabilities greater than 0.1, greater than 0.5, and greater than 0.9. Moving from the lowest to highest cut point, sensitivity decreased from 0.891 to 0.130 as specificity increased from 0.711 to 0.998. New York State's risk adjustment model for predicting death following CABG surgery had a c-statistic of 0.79 (Hannan et al. 1994, 763). As described in Chapter 8, to examine the quality of surgical care in Veterans Administration (VA) hospitals, Khuri and colleagues (1997) developed multivariable logistic regression models to predict 30-day postoperative mortality, using data from over 87,000 noncardiac operations. Of 67 preoperative variables collected on each patient, 44 were statistically significant. However, the first 10 variables entered

in the model contributed most of the explanatory power ($c = 0.87$ for the model with 10 variables versus $c = 0.89$ for the model with 44 variables). Preoperative serum albumin level was the most important predictor; a model based on serum albumin alone achieved a c value of 0.79.

In 2007 CMS began to publicly report hospital-specific risk-adjusted 30-day mortality rates for AMI and heart failure; in 2008 it added pneumonia (see chapters 9, 12, and 17). The model used for risk adjustment for AMI was derived using administrative data from a sample of over 140,000 cases from over 4,660 hospitals and included 27 variables (Krumholz et al. 2006a). Using the development data set, the model had a c-statistic of 0.71. Observed mortality was 4 percent in the lowest decile of predicted risk and 40 percent in the highest decile. Using the validation data, the model had a c-statistic of 0.69. A model using medical record information and developed on the validation data had a c-statistic of 0.77. The heart failure risk adjustment model, which had 24 variables, was based on data from over 222,000 patients from over 5,100 hospitals (Krumholz et al. 2006b). Using the validation data, this model had a c-statistic of 0.70. The model developed using medical record information had a c-statistic of 0.78. In 2009, CMS began reporting 30-day risk-adjusted readmission rates for AMI, heart failure, and pneumonia for each hospital. The risk adjustment model used for patients with heart failure was derived using administrative data from a sample of over 280,000 patients from over 4,660 hospitals. It had 37 variables and a c-statistic of 0.60 (Keenan et al. 2008).

Hosmer and Lemeshow (2000) introduced widely used language judging different c values: Discrimination is acceptable if $0.7 < c \leq 0.8$, excellent if $0.8 < c \leq 0.9$, and outstanding if $c > 0.9$. From this perspective, the CMS mortality prediction models are at best barely acceptable and the readmission models perhaps not even acceptable. However, models' acceptability must be viewed within the context of their use: To whom are they acceptable, and for what purpose? Of note, the National Quality Forum endorsed the CMS models for use in public reporting of mortality and readmission rates, largely because estimates of hospital performance are similar when obtained from either these models or from "gold-standard" models that use medical record data. As Keenan and colleagues (2008, 34) observed, the models are used "to profile hospital performance on the basis of patient status at admission, not to . . . predict outcomes for individual patients."

Researchers have explored the value of adding laboratory values to standard administrative data by examining changes in the c-statistic. Tabak, Johannes, and Silber (2007) developed risk adjustment models for predicting in-hospital mortality for six conditions using data from almost 195,000 patients from 71 hospitals. Risk factors included demographics, admission-based laboratory values, ICD-9-CM variables, vital signs, and altered mental states. Model c-statistics ranged from 0.81 to 0.89. For most of the conditions,

among the various kinds of information available, laboratory data were the most powerful predictors of in-hospital mortality. Escobar and colleagues (2008) developed a model to predict inpatient and 30-day mortality across a wide range of conditions, using data from almost 260,000 patients from 17 hospitals that compose the Northern California Kaiser Permanente Medical Care Program. Predictor variables included demographics, admission type, admission diagnosis, laboratory-based acute physiology score (based on 14 laboratory tests obtained in the 24 hours preceding hospitalization), and comorbidity point score. Using the validation data, the models for predicting inpatient and 30-day mortality had c-statistics of 0.88 and 0.86, respectively. Physiological data accounted for a substantial proportion of each model's predictive ability.

Pine and colleagues (2007) analyzed changes in the c-statistic when adding various kinds of information to standard administrative data. Across several conditions, the mean c-statistics increased from 0.79 (when basic administrative data were used) to 0.84 (after present on admission [POA] codes were added) to 0.86 (after laboratory values also were added) to 0.88 (after supplemental clinical information was added to all of the previous data). POA flags offer important benefits, as described in Chapter 5. Without such codes, it could be problematic to use diagnostic conditions likely to have arisen during the hospital stay as predictors in the risk adjustment model; when coded credibly, only conditions flagged as POA can be used.

Integrated Discrimination Improvement

As illustrated by Pine and collaborators (2007), in some situations additional predictors noticeably increase the c-statistic. However, changes in the c-statistic may not be the best way to measure model improvements. For models that have reasonably high ability to discriminate (e.g., the Framingham risk score, which predicts the risk of cardiovascular disease), potential new risk factors must have very large independent associations with the outcome to meaningfully increase the c-statistic (Pepe et al. 2004; Greenland and O'Malley 2005; Ware 2006). None of the new risk factors proposed to be added to the Framingham model comes close to the necessary levels of association (Pencina et al. 2008). Furthermore, modest changes in c-statistic values (e.g., from $c = 0.84$ to $c = 0.86$) are difficult to interpret.

In the following discussion, we describe *integrated discrimination improvement (IDI)*, proposed by Pencina and colleagues (2008) to help measure the improvement of a new model over a standard model.[15] We illustrate this measure in the context of exploring the benefit of adding laboratory values to demographic and diagnostic data to predict mortality following hip fracture. The data come from 4,344 patients (270 of whom died) treated for hip fracture in 160 VA hospitals over the period 2004 to 2007.

The *IDI* considers changes in an individual predicted probability of an event when one model rather than another is used for prediction. Pencina and coauthors (2008) define the *IDI* in terms of average sensitivity (*IS*) and average "one minus specificity" (*IP*). *IS* and *IP* are similar to the two axes used to define the area under the ROC curve (Exhibit 10.7). We can think of *IS* as the average true positive rate and *IP* as the average false positive rate. The improvement in the model is calculated as

$$IDI = (IS_{new} - IS_{old}) - (IP_{new} - IP_{old}).$$

Thus, *IDI* rewards the new model for improvements in the true positive rate while penalizing it for increases in the false positive rate. If the event of interest is the death of a patient, and

- $ave(\hat{p}_{new,event})$ = mean of the *new model* predicted probabilities of those who die,
- $ave(\hat{p}_{old,event})$ = mean of the *existing model* predicted probabilities of those who die,
- $ave(\hat{p}_{new,noevent})$ = mean of the *new model* predicted probabilities of those who live, and
- $ave(\hat{p}_{old,noevent})$ = mean of the *existing model* predicted probabilities of those who live, then:

$$\widehat{IDI} = [ave(\hat{p}_{new,event} - \hat{p}_{old,event})] - [ave(\hat{p}_{new,noevent} - \hat{p}_{old,noevent}).$$

As noted earlier, to examine model discrimination, Yates (1982) suggested plotting side-by-side histograms of predicted probabilities for those who live and those who die. The difference in the means of these two sets of probabilities, which is approximately equal to R^2 (exactly equal if the model is fit using OLS, as we discuss later), is sometimes called the *discrimination slope* (*DS*) and sometimes the *coefficient of discrimination* (Tjur 2009). Rearranging terms shows that the *IDI* can be expressed as the improvement in discrimination slope between the old and new models:

$$\widehat{IDI} = DS_{new\ model} - DS_{old\ model}.$$

This measure also approximately equals the difference in R^2 between the two models.

The standard error of each component of $\widehat{IDI}$ (i.e., one involving those who experience the event and one involving those who do not) is calculated similarly to the standard error in a paired *t*-test, as follows: (1) Calculate the difference in probabilities using the new and old model, (2) calculate the standard deviation of the differences, and (3) divide the standard deviation of the

differences by the square root of the number of cases who experience or do not experience the event. If these standard errors are denoted as $\widehat{SE}_{event}$ and $\widehat{SE}_{noevent}$, the z-statistic for the null hypothesis that $IDI = 0$ is

$$z = \frac{\widehat{IDI}}{\sqrt{(\widehat{SE}_{event})^2 + (\widehat{SE}_{noevent})^2}}.$$

In our example,

$$ave(\hat{p}_{new,event}) = 0.112 \text{ and } ave(\hat{p}_{old,event}) = 0.096, \text{ and}$$
$$ave(\hat{p}_{new,noevent}) = 0.053 \text{ and } ave(\hat{p}_{old,noevent}) = 0.053.$$

Thus, $\widehat{IDI} = 0.016$ (and $z = 3.62$). From the component analysis of the IDI, we conclude that when the new model is used, sensitivity (true positive rate) increases by 0.016, or almost 17 percent ($0.016/0.096 \times 100$), with no concomitant increase in "one minus specificity" (false positive rate). Again, this analysis provides a better sense of the impact of the new model than does the change in the c-statistic.

Note that if the old and new models have been developed on a data set different from that on which they are evaluated, the models may not calibrate well, which throws off the discrimination slope. One way to address this problem is to recalibrate the predicted probabilities of events by multiplying them by the ratio of the observed event rate to the mean predicted probability of an event. This recalibration ensures equality between the observed and predicted event rates in the data set on which the models will be evaluated.

Methods for Measuring Model Calibration

The calibration curve graphically compares predicted probabilities of a dichotomous outcome with actual outcomes (Yates 1982). To construct it, we divide patients into groups based on similar predicted probabilities of death (e.g., deciles of predicted probabilities of death). The average of the predicted probabilities for patients in each group is plotted against the rate of actual outcomes in the group. Both Escobar and colleagues (2008) and Tabak and collaborators (2007) showed calibration curves for their mortality prediction models.

Lemeshow and Hosmer (1982) proposed a χ^2 (chi-square) test derived from data organized similarly to the calibration curve. The Hosmer-Lemeshow χ^2 is often used in evaluations of risk adjusters to examine model calibration across the range of predicted probabilities. The statistic does not check directly for overall calibration (i.e., whether the average of predicted outcomes is close to the actual outcome rate). Instead, it addresses whether

average and predicted rates are similar within population subgroups. Although the Hosmer-Lemeshow method can be performed in several ways, the following four steps are an increasingly standard application (in this case, death is the target outcome):

1. Data are divided into ten subgroups based on deciles of predicted risk (*PRED*) of death.
2. Within each subgroup, deviations between observed and expected numbers of deaths (O_{die} and E_{die}) and observed and expected numbers of live cases (O_{live} and E_{live}) are measured using a statistic similar to χ^2 (Lemeshow and Hosmer 1982):

$$\frac{(O_{die} - E_{die})^2}{E_{die}} + \frac{(O_{live} - E_{live})^2}{E_{live}} = \frac{(O_{die} - E_{die})^2}{npq},$$

   where n = the number of cases in the subgroup, p = the predicted death rate for these n cases, and $q = 1 - p$. The expected numbers are determined from the *PRED*s.
3. To assess whether deviations are larger than expected (under the hypothesis that each prediction is correct), we sum these deviations over the ten subgroups. The result is compared to a χ^2 distribution with eight df (Lemeshow and Hosmer 1982).
4. The model is accepted if the p-value for this test is reasonably large (i.e., the observed deviations from model predictions are consistent with the typical size of deviations that would occur by chance if the model were correct).

As for any χ^2 tests, a serious concern when using this or any version of the Hosmer-Lemeshow approach is that the decision about the adequacy of model fit depends heavily on the number of observations available. When the sample size is large, even small discrepancies between the model's predictions and observed counts (which may have little practical significance) can produce low p-values, causing rejection of the model. In contrast, when few cases are available, the null hypothesis (that the model is correct) may be accepted despite large differences between expected and observed values.

Other versions of this test use different rules for assigning cases to subgroups or strata. With fewer cases, for example, analysts could split the data into five quintiles of predicted risk rather than ten deciles. When large fractions of the cases have very similar predicted probabilities, cases have to be grouped by a different strategy altogether. For example, if the *PRED*s for the healthiest half of the population are all similarly low, the lowest subgroup might contain the 50 percent of cases with the lowest *PRED* values.

Strata do not need to contain equal numbers of cases, but analysts should have a clear and plausible rationale for their grouping rule. They should also show the distribution of cases among the subgroups. With k strata, the χ^2 with $k - 2$ df is the test statistic. The outcome of the test (i.e., the decision about whether the model is calibrated) may depend on the rule for grouping patients into strata.

Another concern with the Hosmer-Lemeshow χ^2 test is that npq overestimates the variance of the deviations under the null hypothesis. If subgroups are formed using deciles of predicted risk, npq and the sum of the p_iq_is are generally similar, causing little difficulty. However, when the subgroups contain cases with widely dispersed probabilities of death, the Hosmer-Lemeshow statistic is noticeably smaller than it should be, leading to the inappropriate acceptance of models whose predictions deviate greatly from actual outcomes.[16]

For the New York State CABG model, the p-value for the Hosmer-Lemeshow χ^2 statistic was 0.16, indicating good fit, particularly in light of having over 57,000 cases in the database (Hannan et al. 1994, 763). The Veterans Health Administration's models for predicting 30-day mortality for noncardiac surgery cases were developed on over 87,000 cases. They did not calibrate well, though models for specific types of surgery did: "The only goodness-of-fit statistics that were statistically significant at the 0.05 level were for all operations combined and for general surgery, primarily because there were a large number of cases in these categories" (Khuri et al. 1997, 320).

In summary, the Hosmer-Lemeshow test has drawbacks, especially its sensitivity to the number of cases studied and the rule for assigning cases to subgroups. We therefore suggest that analysts treat Hosmer-Lemeshow test results as exploratory. Differences in actual outcomes across the strata may provide more insight about the value of a model (by illustrating how well it discriminates) than do comparisons of actual and observed outcomes within strata (to test how well it is calibrated).

Model Performance Within Subgroups

The Hosmer-Lemeshow statistic provides a single summary measure of the match between predicted and actual outcome rates within subgroups of the data. Nevertheless, even with poor summary fit, analysts can look individually at each pair of observed and predicted outcome rates to see which deciles contribute most to the poor fit and to provide additional insight into model performance. Ash and colleagues (1990) used a deciles-of-risk calibration table to illustrate the ability of a model developed on administrative data available at the time of hospital admission to predict whether utilization reviews would deem the admission inappropriate. Using a model fit to the whole data set (after examining models fit to two halves of a split sample),

they examined calibration in deciles by comparing figures within rows, as shown in Exhibit 10.10. Except for deciles 5, 9, and possibly 7, the observed and predicted problem rates are similar. The final column, which records each decile's contribution to the Hosmer-Lemeshow statistic, confirms that the other rows are not widely discordant. For the Hosmer-Lemeshow statistic not to reject the model fit, the sum of the numbers in the last column must be less than 15.5, which is the 95th percentile for a χ^2 distribution with eight df. The numbers in Exhibit 10.10 add to 11.9, so the model passes this test of calibration.

Examination of the observed percentage problems column (i.e., the observed problem rate in each decile) of Exhibit 10.10 provides further insight into model discrimination. If the model had no explanatory power, approximately 8.3 percent of cases in each decile would have the outcome. However, the outcome rate was dramatically higher for cases in the top decile than for other cases, and few problem cases occurred in the lowest four deciles. Cases in deciles 5 through 9 had outcome rates relatively close to the outcome rate of 8.3 percent for the whole sample. Ash and Byrne-Logan (1998) used an analogous methodology to assess model performance when predicting a continuous outcome, next year's cost.

EXHIBIT 10.10
A Deciles-of-Risk Calibration Table

Decile	Predicted % Problems	Observed % Problems	Contribution to the Hosmer-Lemeshow Statistic
1	0.3	0.1	1.3
2	1.4	1.6	0.2
3	2.7	2.3	0.6
4	4.0	4.1	0.0
5	6.1	7.5	3.6
6	6.9	6.5	0.3
7	7.8	9.0	2.3
8	8.7	8.1	0.5
9	11.4	9.7	3.0
10	34.1	34.5	0.1

Note: The contribution to the Hosmer-Lemeshow statistic for decile i is $(O_i - E_i)^2/n_i p_i q_i$, where O_i is the observed number of problem cases in decile i, E_i is the expected number of problem cases in decile i, n_i is the number of cases in decile i, p_i is aver (*PRED*) in decile i, and q_i is 1 – aver (*PRED*) in decile i. The number of cases in decile i (n_i) is always within one of 1,030.5, since N = 10,305 is the full population size. The Hosmer-Lemeshow statistic is 11.9, the sum of the numbers in the last column. The probability that a χ^2 statistic with 8 degrees of freedom is greater than 11.9 exceeds 10 percent. Thus, the model is not rejected, $p > 0.10$. Observed problem rate = 8.334 percent. Predicted problem rate = 8.336 percent.

We can compute the fraction of all outcomes that occur in a decile or a group of deciles by adding the observed percentage of outcomes for these deciles and dividing by the sum of all the percentages (83.3 percent for Exhibit 10.10). This computation addresses useful practical questions, such as: What percentage of all complications was found by looking at the 10 percent of the cases thought to be at greatest risk of complications? If we use the data from Exhibit 10.10, the answer is over 40 percent: 34.5/83.34 = 0.41. In contrast, fewer than 10 percent of outcome events occurred in the bottom four risk deciles.

Hannan and colleagues (1994, 765) used the Hosmer-Lemeshow statistic computed in deciles of predicted risk to examine the performance of New York State's CABG risk adjustment model, concluding "there was a reasonably good correspondence between actual deaths and expected deaths in each of the individual intervals." They noted that in the highest decile, predicted mortality was 36 percent and actual mortality was 32 percent; thus, the model appeared to adjust sufficiently for the most severe cases.

Similar tables can be constructed for other divisions of the data into subgroups (such as by age or payer class) to identify types of cases whose rate of events is not predicted well by the model. Analysts must then determine why predictions are poor for these cases.

Both calibration and discrimination are important to evaluating model performance. However, Lemeshow and Hosmer (1982) argued that if models do not calibrate well, there is little value to examining their ability to discriminate. Harrell and colleagues (1984, 144) voiced another view: "The reason we argue for first priority to discrimination in judging a model's relative performance is that if discrimination deteriorates, no adjustment or calibration can correct the model. On the other hand with good discrimination, one can calibrate the predictor to attain reliability without sacrificing discrimination."

If the model is used to distinguish those with versus without the outcome, actual values of the *PRED*s are not important. However, if the model is used to determine an expected outcome rate for comparison with an actual outcome rate (e.g., as in profiling hospitals or physicians on the basis of their risk-adjusted mortality rates), calibration is more important than discrimination.

R^2 and Dichotomous Outcomes

As for continuous outcome variables, R^2 for models predicting dichotomous outcomes can be defined several different ways (Tjur 2009). We use the same definition as that for continuous outcomes:

$$R^2 = 1 - \frac{\sum_i (Y_i - PRED_i)^2}{\sum_i (Y_i - \overline{Y})^2}.$$

The predicted probabilities ($PRED_i$s) can come from any method of modeling, including OLS, logistic regression, or a binary splits algorithm. If a model targeting mortality is fit to data using OLS, R^2 exactly equals the amount by which the average predicted probability of death among those who died (AVE_1) exceeds the average predicted probability of death among those who lived (AVE_0).[17] When predictions are derived using a logistic regression model, R^2 usually closely approximates this difference.

The same issues associated with using R^2 to measure performance for modeling continuous outcomes (discussed earlier) arise when modeling dichotomous outcomes. However, for dichotomous outcomes modeled using techniques other than OLS, some question the value of R^2. Ash and Shwartz (1999) show that R^2 provides real information about the accuracy of models to predict 0/1 outcomes that the c-statistic does not provide.

All general warnings about difficulties interpreting summary measures of model performance apply whether R^2, c-statistics, or some other measure is reported. In particular, performance measures calculated on different data sets may not be comparable, especially if the populations vary markedly in pertinent characteristics or in the amount of total variability the model attempts to explain. As discussed earlier, this problem arises because the level of performance achieved by a model on a particular data set largely depends on the variance of important predictor variables and of the outcome. Many researchers fail to appreciate that this limitation pertains to the c-statistic as well as to R^2 and other measures.

Hadorn and colleagues (1992) explored the effect of the "hardness" (i.e., difficulty) of the data set on performance measures when modeling a dichotomous outcome. On the basis of predictions from a logistic regression model developed to predict mortality within 30 days of admission for patients with AMI, the overall data were divided into subsets that differed in the dispersion of the predicted probabilities. They ran both good and poor models on the harder and easier subsets of the data. Measures of poor models' performance on easy data often exceeded measures of good models' performance on difficult data, demonstrating the importance of comparing models on the same data (Hadorn et al. 1992). The problem of noncomparability can be partially addressed with a weighted analysis, but such analysis is possible only if discrepancies between data sets result from well-documented and objectively quantified reasons (e.g., differences in the way in which cases were sampled within strata). While dissimilarity in the outcomes of two data sets clearly suggests noncomparability, similarity does not guarantee that the two data sets are equally easy to model (Hadorn et al. 1992).

As described in the following section, cross-validated R^2 values for dichotomous outcomes have ranged generally from 0.10 to 0.30 in various studies of risk-adjusted death rates. Typically, models with higher c-statistics have higher R^2 values, although exceptions do occur. Two models applied to

a single data set often have similar c-statistics but different R^2 values. Indeed, c-statistics may be less sensitive than R^2 to real differences between models. To illustrate, we fit two models to predict in-hospital death for 17,577 AMI patients from the MedisGroups 1991 Comparative Database (Iezzoni et al. 1994b). These data included information from 112 hospitals nationwide that used the MedisGroups risk adjustment method. The in-hospital death rate was 13.5 percent. Model 1 included dummy variables for the five original MedisGroups severity classes. Using administrative data only, model 2 included 8 dummy variables for age categories and 12 dummy variables to represent the presence of 12 chronic conditions defined using discharge ICD-9-CM diagnosis codes.

For model 1, $R^2 = 0.15$ and $c = 0.76$; for model 2, $R^2 = 0.07$ and $c = 0.72$ (Iezzoni et al. 1994b, 45). Therefore, both R^2 and c suggested that model 1 performed better than model 2. We used bootstrapping (see discussion later in this chapter) to determine empirically the distributions of, and 95 percent confidence intervals around, c and R^2 for the two models. Because of the large sample size, these confidence intervals were narrow, and differences in performance between the two models appeared statistically significant. But the superiority of model 1 was more striking when we looked at R^2 versus the c-statistic (R^2 for model 1 was 0.15 versus 0.07 for model 2, as opposed to c-statistics equaling 0.76 and 0.72, respectively). We standardized each difference by dividing it by a pooled estimate of its standard deviation. The two resultant R^2 values differed by ten standard deviations, while the two c-statistics differed by only six. If we had considered a smaller sample size and only c-statistics, we might not have detected the superiority of model 1. Thus, R^2 may make differences more evident than do c-statistics (Ash and Shwartz 1999).

The bottom line is that both c and R^2 are useful measures of the performance of models predicting dichotomous outcomes. With binary outcomes, however, many analysts do not appreciate how powerful models with small R^2 values can be. Thus, reporting only R^2 may expose a good model to criticism. In general, attempts to argue "policy value" require descriptions of model performance that are easier to grasp than R^2 and c. A table of actual death rates within deciles of predicted risk (e.g., Exhibit 10.10) is often helpful in this context.

Model Validation

Models fit to specific data sets must undergo validation to determine how well they might perform in other settings (see Chapter 9). Analysts typically validate models through two approaches: cross-validation, in which part of the same data set used to develop the model is used to validate the model, and independent validation, in which the model is applied to entirely new data.

Because analysts frequently do not have access to new databases, they commonly use a portion (e.g., one-half or two-thirds) of their data to develop the model and the remaining portion (e.g., one-half or one-third) to validate the model. Sometimes researchers have access to additional years of data, which can be used for validation under the assumption that relationships are not changing over time. For example, Krumholz and colleagues (2006a) developed a risk adjustment model for AMI based on administrative data. CMS uses this model to risk-adjust hospital-level mortality rates and then posts them on the CMS Hospital Compare website (see Chapter 12). Approximately one-half of the sample of about 280,000 cases in 1998 was used for model development, and the other half was used for model validation. In addition, they used data from other years (1995 to 2001) and data that could be linked to medical chart reviews (about 180,000 cases from 1994 to 1995) as additional validation samples. Models developed from administrative data had comparable c-statistics (approximately 0.70) for the development and validation data sets. The model developed from chart review data performed better (c-statistic = 0.77). However, the correlation between standardized mortality rates derived from the administrative model and from the chart-based model was high (0.90).

In our study of hospital-based severity measures, we had only one year of data, and we performed cross-validation as follows. First, we randomly split the data in half, into samples 1 and 2, and fit models to each sample. In sample 1, we computed SSE_1 and SST_1 as

- SSE_1 = sum (within sample 1) of squared differences between each Y and the $PRED$ predicted for it by the model that was fit to sample 2, and
- SST_1 = sum (within sample 1) of squared differences between the individual Ys and the average Y in sample 1.

Second, we computed a cross-validated R^2, denoted CVR^2_1, as $1 - (SSE_1/SST_1)$ and then repeated this process, reversing the roles of samples 1 and 2, to find CVR^2_2, a validated R^2 for the sample 2 data. We averaged these two values to compute a summary CVR^2. This method provided insight into the variation associated with the resulting value. Another approach is to calculate a single CVR^2 value on the whole database directly, using

- $SSE = SSE_1 + SSE_2$ = sum (over all cases) of squared differences between each Y and the $PRED$ predicted for it by the model that was fit to the sample that did not contain it, and
- SST = sum (over all cases) of squared differences between the individual Ys and the average Y in the whole population.

Note that SST does not equal $SST_1 + SST_2$.

This approach offers the advantage of generalizing readily to settings in which half of the data are viewed as inadequate for model development.[18] In this situation, a popular cross-validation approach is to divide the data into a number of approximately equal parts—for example, ten—and then fit models successively to all of the data except one part. Each model developed on 90 percent of the data is then used to make predictions for the remaining 10 percent. Thus, the prediction for each case is based on a model that did not use that case for parameter estimation. Cross-validated R^2 values are then calculated in the usual way using the actual and predicted values for all cases.

Knaus and colleagues (1993) held out successive tenths of cases in estimating the cross-validated R^2 for predictions of length of stay using APACHE III. In creating the Medicare Mortality Predictor System (see Chapter 8), Daley and colleagues (1988) used successive subsamples of 90 percent of the data to make predictions for the holdout 10 percent subsamples; they calculated R^2 values from the predictions for the holdout subsamples. To validate the VA mortality prediction models, Khuri and collaborators (1997) divided the data into halves and then developed their models on half the patients and tested them on the other half. For the model for all types of surgeries, they repeated this process three times, comparing the c-statistics for the fitting and validation sample each time. On average, the c-statistic was 0.003 lower for the validation sample. For models developed for specific operations (which had much smaller sample sizes), the c-statistic for the validation sample generally fell within 0.05 of the c-statistic for the fitting sample. However, differences of that magnitude (0.05) in c are not trivial. Typically, models with c-statistics of 0.70 are notably more accurate than those with c values of 0.65.

In general, the smaller the holdout sample, the greater the concern about falsely assessing model performance. If each iteration holds out only small subsets of cases, the remaining main body of data is similar for each successive model fit (Picard and Berk 1990). Thus, the models are unlikely to differ much from a model fit to the entire database. Nevertheless, cross-validated R^2 values can differ substantially from R^2 values for the development data set, even when only one case is held out.

It is a good idea when comparing R^2s and c-statistics for different models to recognize that these measures are statistics that could vary randomly from sample to sample. *Bootstrapping* enables analysts to estimate confidence intervals for statistics whose distributional properties are not well characterized (see Chapter 8). To construct a bootstrap estimate of the confidence interval for a model's R^2, analysts sample with replacement from the original data set (i.e., each item is put back in the sample before a new item is drawn) to create a large number (perhaps 1,000) of simulated data sets, each the same size as the original data set.[19] The model is fit to each simulated data set, and R^2 is calculated. Confidence intervals for R^2 are constructed from the distribution of

the 1,000 simulated values. For example, if the 5th and 95th percentiles (i.e., the 50th from the lowest and 50th from the highest of the 1,000 ranked R^2 values) were 0.15 and 0.23, respectively, then [0.15, 0.23], or 15–23 percent, would be a 90 percent confidence interval for R^2.

In assessing the contributions of severity of illness to explaining cost differences between tertiary teaching and nonteaching hospitals, we estimated that severity plus other patient characteristics (e.g., admission source, discharge destination, transfer status, purpose of admission) explained 18 percent of the higher costs at teaching hospitals. We used bootstrapping to determine a 90 percent confidence interval (4–33 percent) for this estimate (Iezzoni et al. 1990). Khuri and colleagues (1997) used bootstrapping to examine the stability of the important risk factors entered in stepwise logistic regression models to predict postoperative deaths. For each type of surgery, they drew 15 bootstrap samples and developed models. They measured the stability of each variable by the percentage of the models in which a given variable appeared. Serum albumin, the most important predictor, appeared in 92 percent of the replications, followed by age (84 percent), emergency status (84 percent), American Society of Anesthesia classification (76 percent), and disseminated cancer (74 percent).

Confidence intervals for R^2 and model stability are usually estimated from bootstrap replications, as are confidence intervals for c-statistics. However, Hanley and McNeil (1982, 1983) provide methods for analytically calculating confidence intervals for c-statistics and differences in c-statistics.

The strongest form of model validation is application of the model to a new data set. However, if a model does not validate well, the model is not necessarily the problem. As noted earlier, when model parameters are estimated on a development data set using OLS, the average predicted values from the model $AVE(PRED_i)$ equal the average actual values of the dependent variable $AVE(Y_i)$. When maximum likelihood methods are used to estimate parameters (e.g., in a logistic regression), the averages are usually close but not necessarily equal. However, when a previously developed model is applied to new data, $AVE(PRED_i)$ generally differs from $AVE(Y_i)$. As discussed earlier, the significance of this difference depends on the purpose of the comparison. If the goal is to assess a model's validity, the difference between $AVE(PRED_i)$ and $AVE(Y_i)$ measures model calibration. Large differences suggest poor validation. Alternatively, "miscalibration" of the model may indicate that the two settings from which the independent databases come differ in important ways, such as in their quality or efficiency of care. No foolproof way exists to choose between these competing hypotheses (poor validation versus setting differences)—one reason why examining trends with a longitudinal data set can be so powerful (see Chapter 12). From year to year, data from the same setting are generally more comparable than are data from completely independent sources.

As discussed, summary measures of model performance are likely to deteriorate when models are applied to independent data. The most complex models experience the greatest loss. However, such deterioration does not necessarily occur. For example, Ash and colleagues (2003) used data from 1995 on more than 300,000 cases to estimate parameters to predict one-year mortality following AMI using three models: a Charlson-like comorbidity model including 17 conditions, a DCG model including 118 conditions, and a model including the 259 conditions in the Agency for Healthcare Research and Quality's comorbidity software. They then used the models to predict mortality in 1999. The c-statistics from the models fit to 1995 data were 0.73, 0.80, and 0.81, respectively. Given the large number of cases involved, the validated c-statistics were also high. However, because of the numerous parameters, some deterioration in the c-statistic for the DCG model, and even more deterioration in the c-statistic for the model including 259 conditions, was expected. Surprisingly, in all cases, c-statistics increased by 0.01 for the 1999 validation data. The likely explanation is twofold: The 1995 data were sufficiently large to avoid overfitting, and the newer data were somehow easier to fit, perhaps because of better diagnosis coding.

In addition to evaluating validity with overall summary measures, assessing performance for subgroups of interest using both development and validation data sets is valuable. Suppose that a model developed to predict death is applied to some subgroup of cases, such as to a subset of the model development data set or to new cases that share a particular characteristic (as opposed to a randomly sampled validation data set). Suppose further that the observed death rate in the study group is significantly different from the rate the model predicts. If we believe that the model fairly captures and accounts for all important risk factors, we must look for some other causal factor, such as lower- or higher-quality care. Systematic mispredictions for subgroups of patients may also reflect problems with the model.

Blumberg (1991) viewed poor model fit for specific patient subgroups as "biased estimation." To demonstrate the problem, he employed an old version of the admission MedisGroups severity score to adjust for 30-day mortality of AMI patients and found that estimates of expected mortality differed significantly from actual outcomes in subgroups defined by such attributes as patient age, location of infarction in the myocardium, history of congestive heart failure, serum potassium level, blood urea nitrogen, pulse rate, and blood pressure. For instance, Blumberg reported 38.4 deaths among 289 patients aged 85 or older versus an expected 27.2 deaths. He concluded that this version of MedisGroups was biased by not accounting properly for the independent effect of age on mortality risk. Subsequent versions of MedisGroups specifically included age in the prediction models

(Steen et al. 1993). Although age can be an independent risk factor for mortality (see Chapter 3), Blumberg's finding is consistent with the possibility that very old persons receive worse quality of care, resulting in more deaths than expected.

In developing APACHE IV, Zimmerman and colleagues (2006) were motivated in part by the poor model fit of APACHE III for certain groups of patients. To mitigate this problem, they included more sophisticated modeling of the relationship between age (and a number of other continuous variables) and mortality in APACHE IV through the use of cubic splines and considered a large number of ICU admission diagnoses (430, of which 116 were included in the model). Improved prediction for specific subgroups of patients "support the importance of more precise adjustment for ICU admission diagnosis in prognostic models" (Zimmerman et al. 2006, 1304).

When models are validated on subgroups of the data, problems might arise due to intrinsic differences in the hardness of the data for modeling: Performance measures for models applied to different subgroups are not comparable. Consider a data set that contains patients representing three different diseases. If models for predicting mortality are fit separately to each disease subgroup, R^2 for making predictions for the whole data set is a sum of two quantities: the average of the three R^2 values for the three subsets and R^2 for a model that assigns the average death rate for each disease to each patient with that disease. Thus, if the three diseases have very different death rates, the overall R^2 could be large simply because disease itself is highly predictive, even if the model's ability to distinguish risk among patients with the same disease is negligible.

One should distinguish subgroups formed *a priori* (using variables similar to candidate risk factors) from subgroups based on observed outcomes (e.g., particularly high or low values of the outcome variable). Risk adjustment methods may fail to calibrate well in extreme subgroups formed from observed outcomes, a problem sometimes called *compression*. However, when groups are formed *post hoc* on the basis of extreme outcomes, the mean expected value of the group **should** be less extreme than the actual group value. This problem is especially evident when one considers models for predicting the total cost of next year's health services for a group of persons. For example, nearly 20 percent of Medicare beneficiaries generate essentially no health care costs over one year. However, no Medicare recipient generates an expected cost of $0 during a year of enrollment. No matter how healthy a sizeable group of people was in a particular year, some expenditures inevitably occur in the next year. Similarly, even if a group of persons has high expected costs next year because of problems observed this year, the group's actual costs will likely be lower than the highest Medicare costs next year. Patients with the very highest or lowest costs in any one period

are expected to have costs that are closer to the overall average in the next period, a phenomenon called *regression to the mean* (Bailar and Mosteller 1986, 87).

Many subgroups of policy interest are defined on the basis of characteristics other than the targeted outcomes, such as urban versus rural location, public versus private hospital, Medicaid versus commercial insurer, and insured versus uninsured person. For such groups, discrepancies between observed and expected outcomes must be interpreted carefully. For example, suppose that death rates in public hospitals are significantly higher than expected after risk adjustment based on a particular set of variables. Whether this difference indicates that care in such hospitals is poor or that patients seen in public hospitals are sicker than patients in other hospitals in ways not measured by the model is critical to understanding this finding. Often such results prompt pointed disputes about what exactly the risk-adjusted findings mean and flag a need for further research.

Independent validation can extend beyond statistical assessment of the model to consider the purposes for examining risk-adjusted outcomes. As illustrated in the National Veterans Affairs Surgical Risk Study (see Chapter 8), researchers aimed to answer the question: To what extent do model-based risk-adjusted measures of quality correspond to quality measured using different approaches? Daley and colleagues (1997a) described site visits to 20 VA surgical services with either lower- or higher-than-expected risk-adjusted mortality or morbidity rates. Site visitors rated overall quality and technology/equipment much more favorably at the hospitals with lower-than-expected mortality and morbidity rates. Site visit teams blinded to the predicted outcomes of the hospital correctly determined the outlier status (i.e., higher or lower than expected) for 17 of the 20 surgical services visited. This study powerfully validates the use of the VA risk adjustment model to identify services with high/low quality of care.

Given the strong public desire to compare risk-adjusted outcomes across providers (see chapters 1 and 17), interest in the performance of risk prediction models will increase. Of particular concern is the claim that worse-than-expected outcomes reflect the inadequacy of risk adjustment. The most rigorous way to test this claim is to collect data on the unmeasured risk factors believed to be responsible for the discrepancies and to see how much of the difference is explained. Although theoretically attractive, this test is generally impractical (see Chapter 3). Without further studies of the type conducted by the VA researchers, one must rely on judgments about the face and content validity of the risk adjustment approach (Chapter 9) and judgments about whether unmeasured risk factors might explain observed discrepancies. Thus, interpretation of "objective" measures of performance must ultimately rely on subjective assessments.

Conclusion

A variety of summary measures of model performance exist, including many not discussed in this chapter. However, regardless of what other measures are reported, we believe that all publications and reports about risk adjustment methods should present a core set of common measures. Our short list of important measures includes R^2 and deciles-of-risk tables for both dichotomous and continuous outcomes and, for dichotomous outcomes, the c-statistic and the mean *PRED* for cases with and without the outcome.

Nevertheless, interpretation of these quantitative measures of performance is not straightforward, and subjective judgments are often required. These judgments are influenced by such factors as

- maximum value for the performance measure that could realistically be achieved using the risk adjustment variables on a particular data set;
- specific context or origin of the sample under study (e.g., hospital types, whether certain groups were oversampled);
- variability of the dependent and independent variables in the data set;
- whether and how outliers were identified and handled in the analysis;
- whether the risk adjuster is a continuous variable and, if so, whether and how discrete categories were formed for the analysis;
- whether data transformations were used and, if so, whether measures of model performance used the retransformed data; and
- whether performance measure estimates were validated or confidence intervals were reported.

Notes

1. The buckets could contain equal numbers of cases (quintiles, deciles, or more generally, "quantiles"), capture intuitive cut points (e.g., risks less than 1.0, between 1.0 and 2.0, between 2.0 and 5.0, and over 5.0), distinguish the small but crucial high-risk subpopulation, or match categories that have commonly been used before. No single rule for creating buckets is best.
2. For predicting cost outcomes, 500 or more cases per bucket are desirable.
3. With k buckets, this approach is equivalent to regressing Y on indicator variables for any $k - 1$ of them.
4. Unlike bucketing, multiparameter regressions can produce negative cost predictions for some cases and negative coefficients for factors that logically should not subtract from expected costs. Negative

predictions may be handled in several possible ways; the simplest is to create buckets using cut points of the predicted values and then use as new predictions the average of the actual values in each bucket. Negative coefficients may not be a problem if the major interest is in making good predictions rather than identifying individual effects of specific predictors. However, if desired, negative coefficients on disease indicators can be zeroed out to ensure that coding an additional disease never lowers the risk score.

5. This broad statement is essentially true even when the prediction comes out of a "black box" methodology, such as a neural network.
6. Almost all general-purpose statistical packages include modules that produce locally weighted smooth regressions, such as LOESS.
7. Risk-adjusted payment and risk-adjusted outcomes related to quality assessment assume that patients at higher risk will generate more costs and experience poorer outcomes. As described in Chapter 5, code creep involves assignment of ICD-9-CM codes suggesting more severe conditions than "usual" coding practices would imply. A model that records only the presence/absence of medical conditions, for example, is less sensitive to code creep than is a model that views people as sicker when more codes, or multiple repetitions of the same code, are used to describe essentially one problem.
8. After the hierarchies are imposed on the vector of condition category indicators, the components of the remaining vector are more nearly "orthogonal," which means, for example, that the coefficient associated with simple hypertension is not affected by the outcomes of people who have both hypertension and a heart attack. In a person's HCC profile, hypertension affects only the risk score for people who have hypertension but no more serious heart condition.
9. Here, "number of predictors" is the number of model coefficients or parameters. For example, if x = age is entered using x and x^2, it contributes two predictors.
10. Classification And Regression Trees (CART) was developed by Salford Systems, San Diego, California (www.salford-systems.com). CART is based on techniques described by Breiman and colleagues (1984).
11. One caveat is that it is not clear which, if any, of these distinctions is either statistically (or practically) significant (see discussion of bootstrapped confidence intervals discussed later in this chapter).
12. The central limit theorem states that for large enough n and under very general conditions, the spread of the distribution of a sum (or average) of n independent random variables is a function of *SSE* (not *MAD*). For this reason, R^2 is a preferred measure of model performance. In sufficiently large samples, however, there is a fixed

relationship between the size of *MAD* and *SSE*. Thus, when n is sufficiently large, models with smaller *MAD*s will have larger R^2s.

13. The *SD* of "an individual prediction error," or *SD*(*PRED*), equals $\sqrt{1 - R^2}$ multiplied by the *SD* of *Y*. When the model predicts $\overline{Y}$ for every observation, R^2 is zero and $SD(Y) = SD(PRED)$. Even when a model has an R^2 as high as 20 percent, *SD*(*PRED*) is nearly 90 percent as large as $SD(Y)$ because $\sqrt{1 - 0.20} = 0.894$.
14. Alternatives to creating categories are described in Harrell (2001). Computer-based approaches (e.g., locally weighted regression, i.e., LOESS, or cubic splines) can be used to produce a smoothed plot of the relationship between an outcome and any candidate independent variable, such as a continuous risk measure (see, e.g., Harrell 2001). When creating risk score categories that will be represented in the final model by dummy variables, developers can use change points in a smoothed function (places where the change in outcome is large for a small change in the independent variable) to define "intelligent" cut points.
15. Pencina and colleagues (2008) describe two measures: net reclassification improvement (NRI) and integrated discrimination improvement (IDI). NRI is useful when there is a natural way to categorize risk scores, which is rare for risk scores from the types of risk adjustment models we consider. We discuss IDI, a more useful measure in our context.
16. Let n equal the number of cases in each decile, p equal the average predicted probability of death in the decile, $q = 1 - p$, p_i equal the predicted probability of death of the i^{th} person in the decile, and $q_i = 1 - p_i$. In the Hosmer-Lemeshow statistic, variance in each cell is estimated as npq rather than as the sum of the p_iq_is, although the sum of the p_iq_is is a better estimate. (If the p_is were known to be correct rather than simply estimates, the sum of the p_iq_is would be exactly right.) Furthermore, npq is generally larger than the sum of the p_iq_is. For example, if there are 50 cases with $p_i = 0.90$ and 50 with $p_i = 0.10$, the sum of the p_iq_is equals 9 while $npq = 25$. So, $(O_{die} - E_{die})/npq$ will be only 36 percent as large as would be correct (because $9/25 = 0.36$). On the other hand, if the p_is are split into two groups of 50 with more similar probabilities, say 0.40 and 0.60, the sum of the p_iq_is equals 24, which is 96 percent as large as npq. To avoid this problem, one should use the sum of the p_iq_is instead of npq in calculating the test statistic.
17. The same fact holds for any categorical model (such as would result from using a binary splits algorithm) in which the predicted average in each category is set equal to the actual average.

18. If half the data are inadequate for model development, all the data may still be inadequate because doubling the size of the data set decreases the size of errors by a factor of only $\sqrt{1/2} \sim 0.7$ (~ 30 percent).
19. It might seem counterintuitive to sample N cases with replacements from the data set of N cases that we have. However, the data set we have is our best estimate of the population from which the data arose. What we would really like to do is resample N cases from the population. Because we do not know the population, we sample from our best guess of what the population is: the data set of N cases that we have. In this sense, we are "pulling ourselves up by our bootstraps"—thus the name of the technique.

CHAPTER

11

ESTIMATING THE EFFECT OF AN INTERVENTION FROM OBSERVATIONAL DATA

Michael Shwartz and Arlene S. Ash

Randomized controlled trials (RCTs) are the "gold standard" for evaluating the efficacy of clinical interventions (see Chapter 1). Because RCTs randomly assign patients to intervention and control groups, patients in each group have similar measured and unmeasured baseline risk factors on average. Biases arise in RCTs; for example, study and control subjects may differ with regard to follow-up or the completeness of data capture (e.g., study subjects may be more likely than control subjects to report certain types of symptoms following the intervention even though both groups experience the same effects). Nevertheless, random assignment significantly strengthens confidence that an observed difference in outcomes between the treatment and control groups is due to the intervention rather than baseline group differences.

However, random assignment (basically coin flipping) is rarely used to study most clinical interventions for practical and ethical reasons. For random assignment to treatment or control groups to be ethically justified, genuine uncertainty must exist among clinical experts about the better treatment. In practice, many interventions become widely adopted without adequate study. For example, too little data on women over age 75 are available from RCTs to know whether screening mammography is helpful for this age group. Some believe it would be unethical to withhold this screening test because it has proven benefits for younger women, while others are equally convinced that screening mammography is of little value for older women. Physicians and patients who already believe they know what is best (whichever side they take) often refuse random assignment, making RCTs difficult to conduct.

To increase internal validity, RCTs often recruit relatively homogeneous subjects; for example, they exclude patients over age 75 or those with serious comorbidities. Even among eligible patients, trial participants often differ from those who refuse participation in important ways. Participants are typically healthier than the average eligible person, introducing a "healthy volunteer" effect. In a classic example, perioperative mortality following carotid endarterectomy was higher in actual practice than reported in RCTs, even in institutions participating in the trials (Wennberg et al. 1998). Thus,

findings from RCTs may not generalize well to patients seen in routine clinical practice. Clinicians' belief that study subjects are not like their patients slows the diffusion of RCT-validated interventions. Also, RCTs are expensive and often require years to conduct, during which time the targeted interventions may become obsolete or clinically entrenched, hindering efforts to change practices on the basis of RCT results.

For these reasons, most clinical interventions are not evaluated in RCTs, and analysts therefore must rely on quasi-experimental or nonexperimental observational studies.[1] In observational studies, some patients receive interventions while others do not for reasons that are only partially understood/captured by measurable factors. Analysts compare the two groups' outcomes, hoping to quantify the effectiveness of the intervention. However, direct comparison of outcomes may be misleading: The measured and unmeasured baseline risk factors of persons receiving/not receiving the intervention often differ markedly and thus so do their prognoses. The size and direction of differences are often unknown. Hence, it is difficult to determine the extent to which differences in outcomes are caused by the intervention versus by differences in baseline risk.

When assessing the value of an intervention from observational data, analysts typically use a multivariable model to adjust for baseline differences in risk. Standard multivariable models estimate treatment effectiveness without bias when the following conditions are met: Every important risk factor is measured, no data are missing, and relationships between the risk factors and outcome are correctly specified. However, this conceptual ideal is never attained in real data.

In this chapter, we first briefly discuss standard multivariable models used to make risk-adjusted assessments when the data and models approximate the ideal. We then describe two methods for addressing concerns about the validity of assumptions that underlie the ideal: propensity score matching, which protects against biases due to incorrect model specification, and instrumental variables, which address suspected differences in unmeasured risk factors. Finally, we discuss sensitivity analysis, focusing on the possibility that an unmeasured confounder could have accounted for the observed estimate of intervention effectiveness.

Multivariable Models for Estimating the Effects of Interventions

Continuous Outcomes

Consider estimating the effect of an intervention on a continuous outcome, such as how enrolling in a managed care organization (MCO) influences a person's health care expenditures over a year. The data reflect people enrolled

in MCOs and people who receive their care through a fee-for-service arrangement. Historically, MCO enrollees were healthier than those in fee-for-service. If actual costs for MCO enrollees are lower, how much (if any) of that difference is due to managed care versus differences between MCO enrollees' and fee-for-service patients' clinical needs? The standard approach is to adjust for differences in baseline patient risks using multivariable linear regression.

To generalize, let Y_i be the actual value of the outcome variable for the i^{th} patient and X_{ij} be the value of the j^{th} independent variable for the i^{th} person, $j = 1, \ldots, J$. These J independent variables are the baseline risk factors requiring adjustment. Let I be a variable coded 1 if the patient receives the intervention (is a member of an MCO) and 0 if the patient does not (remains in fee-for-service).

The standard multivariable linear regression model used to assess treatment effectiveness is usually written as

$$E(Y_i) = a + \sum_j b_j X_{ij} + cI,$$

where $E(Y_i)$ is the expected value of the outcome variable for the i^{th} patient, a is a constant coefficient, b_j is the coefficient that measures the systematic effects of the risk factor (X) on the outcome Y, and c is the coefficient that measures the effect of the intervention on Y.

Notice that we use the standard mathematical convention that when two variables are placed adjacent to each other, multiplication is implied. For example, cI stands for $c \times I$. Sometimes, for emphasis, the $\times$ is explicitly written out.

Imagine two patients with the same values for all risk factors, one of whom is in managed care ($I = 1$) and the other in fee-for-service ($I = 0$). Under this model, the cost for a managed care patient is predicted to equal c plus the cost for a fee-for-service person with the same risk factors. If the difference between c and 0 is statistically significant, analysts would conclude that managed care influences cost, in which case they interpret the absolute value of c as the size of the effect of the intervention (MCO membership). For example, if $c = -500$, managed care is associated with a $500 reduction in yearly costs.

How risk factors are coded may be important. For example, if age is coded as a continuous variable, the linear model encodes the constraint that the difference in cost between otherwise comparable 20- and 40-year-olds is the same as the difference in cost between 60- and 80-year-olds. If the anticipated effect of age is nonlinear, age can be entered into the model as a categorical variable—for example, several indicator variables for five-year age categories.[2] "Spline fitting" captures nonlinear relationships between a continuous risk factor and an outcome more elegantly than categorizing does

(Harrell 2001), as do fractional polynomials (Royston, Ambler, and Sauerbrei 1999). However, these approaches are used less commonly, partly because the effect of a categorical variable is easier to convey in tabular format. A reasonable compromise is to use splines (see, e.g., Zimmerman et al. 2006) or fractional polynomials in modeling, but to display the relationships for categories of the risk factors. Royston, Ambler, and Sauerbrei (1999), for example, suggest showing regression-based estimates of the outcome at midpoints of risk factor categories.

An interaction term allows the effect of a variable on the outcome to differ depending on the value of other variables. Interactions between risk factors and the intervention variable are often of particular interest. Consider a model with the following indicator variables: I for the intervention; F for female, coded as 1 for women and 0 for men; and IF (I multiplied by F) for an interaction between gender and treatment, which is 1 for a woman who receives the intervention and 0 for everyone else. The model might look like

$$E(Y_i) = a + \text{the combined effects of other risk factors} + bF + cI + dIF.$$

Thus, regardless of whether interventions occur, women are expected to cost some fixed amount more (or less, if b is negative) than similar men cost.[3] Among men (for whom $F = 0$ and $IF = 0$), the intervention adds c to the expected outcome; among women, the intervention increases the expected outcome by $c + d$. Thus, the interaction term accommodates the possibility that the benefit of the intervention might differ between women and men.

The model parameters (a, the b_js, c, and so on) are estimated from the data to make the set of predicted values, the $PRED$s, or estimated $E(Y_i)$s match the actual values, the Y_is, as closely as possible. The ordinary least squares (OLS) approach is used most often to estimate parameters in multivariable linear regression models (see Chapter 10), though theoretical reasons may support use of more complex methods.[4] To assess the effect of the intervention, we examine the confidence interval and p-value associated with c, the coefficient of the intervention indicator and any other coefficients attached to the intervention, such as d in the model just discussed.

Dichotomous Outcomes

Logistic regression is the standard method for modeling a dichotomous outcome. The dependent variable is the log of the odds of an event of interest (e.g., being alive 30 days after hospital admission). If p_i is the probability of the event for the i^{th} person, then O_i, the odds of the event for that person, is defined as $p_i/(1 - p_i)$.[5] A logistic regression model for assessing treatment effectiveness has the form

$$\ln O_i = \ln\left(\frac{p_i}{1 - p_i}\right) = a + \sum_j b_j X_{ij} + cI,$$

where ln is the natural (base e) logarithm. This model says that the log odds of person i being alive ($\ln O_i$) increases by a fixed amount, c, when the intervention is received. This equation can be rewritten as

$$O_i = e^{(a + \Sigma_j b_j X_{ij} + cI)} = [e^{(a + \Sigma_j b_j X_{ij})}] \times e^{(cI)} = [\textit{person-specific risk}_i] \times e^{(cI)},$$

meaning the effect of the intervention is to multiply a person's odds by e^c.[6] Under this model, dividing the odds of the event for any person receiving the intervention by the odds of the event for the same person when not receiving the intervention always yields e^c. Thus, e^c is the odds ratio (OR)—that is, it is the ratio of the odds for the event among the patients who receive the intervention divided by the odds for the event among patients not receiving the intervention (i.e., the probability that the event happens divided by the probability that it does not). Treatments that make the event more likely have ORs greater than 1, and vice versa. This functional form assumes that the intervention has the same (multiplicative) effect regardless of a person's baseline risk. Here, again, interaction terms could be used to create a more flexible model if this one is too simple to capture important realities.

Logistic regression models estimate model parameters using a maximum likelihood procedure.[7] Issues in specifying the relationships of independent variables to the log odds of the outcome are very similar to those in specifying the relationship of independent variables to a continuous outcome (see previous discussion). As for OLS multiple regression, our particular interest is in the confidence interval and p-value for c.

Time to Event Outcomes

When time to an event (e.g., death) is the outcome, analysts usually must deal with the fact that some cases are not followed until the event is observed; for example, some cannot be located for follow-up, and others are still alive when the study ends. Because these cases were alive on the date they were last seen, we have some idea of how long they survived; however, their exact date of death is censored. The proportional hazards model, discussed briefly in Chapter 10, is standard for modeling survival data with censoring (Cox 1972). Virnig and colleagues (2000) provide a more detailed yet introductory-level discussion of issues in survival modeling, as do Singer and Willett (2003).

The dependent variable in the proportional hazards model is the instantaneous hazard rate at time t, denoted by $h(t)$. This rate can be thought of as the probability that a person who survived to time t died (or experienced the event of interest) in a small time interval from t to $t + \Delta t$.[8] It is expressed as a function of a baseline hazard, $h_0(t)$, treated as a nuisance parameter, and other risk factors. The model is

$$h_i(t) = h_0(t)e^{a + \Sigma_j b_j X_{ij} + cI} = h_0(t)e^{a + \Sigma_j b_j X_{ij}} e^{cI} = [h_0(t)e^{a} + e^{b_1 X_{i1}} e^{b_2 X_{i2}} \ldots] e^{cI}.$$

Implicit in the structure of this model is the assumption that each independent variable affects the baseline hazard by a multiple that is the same for all patients regardless of other characteristics and at all times t. In particular, the effect of the intervention on the hazard rate is found by multiplying by e^c, which is called the hazard ratio (comparable to the OR when the outcome is dichotomous). More complex, time-dependent models are appropriate when the effects of variables change over time (Kalbfleisch and Prentice 2002).

Proportional hazards models use maximum likelihood procedures to produce model parameter estimates, confidence intervals, and p-values, as do logistic regression models. From the hazard rates, analysts can calculate survival functions that predict the probability of surviving at least as long as t units of time, for a range of values of t. Survival functions can be compared for patients with specific characteristics, such as those who do or do not receive an intervention.

Moving Beyond Standard Multivariable Models

The multivariable models just described correctly adjust for differences in baseline risk between intervention and control groups if (1) the important risk factors are observed and measured for all persons and (2) the relationships between risk factors and the outcome of interest are correctly specified. However, measuring every important risk factor for everyone is difficult if not impossible (see Chapter 3), as is ensuring that all relationships are correctly specified. In the next sections, we discuss two options for handling situations in which information about risk factors or about their relationships with outcomes is seriously incomplete or unavailable.

The first option applies when the most important risk factors have been measured but the relationships between risk factors and the outcome may not have been correctly specified. If risk factors are both strongly associated with the outcome and distributed very differently in the intervention and control groups, model misspecification can lead to biased estimates of the intervention's effect. Propensity scores address this problem. The second option applies when important **unmeasured** risk factors that are weakly associated with measured risk factors are distributed differently for persons receiving/not receiving the intervention. This situation is often plausible, as discussed in the next section. We illustrate the use of an instrumental variable to assess the effectiveness of an intervention in such a setting. A number of authors, including Morgan and Winship (2007), place both approaches in the more general framework of estimating causal effects from observational data. An entire issue of *Health Services and Outcomes Research Methodology* (2001, volume 2, issue 3–4) is devoted to causal modeling with health care examples.

Propensity Scores

Propensity scores are useful when

1. several measured risk factors have a strong relationship to an outcome,
2. persons who receive an intervention and those who do not have markedly different distributions of values for these risk factors, and
3. we are unsure about the true relationship between the risk factors and the outcome.

Standard approaches to addressing this situation exist. First, with only a few important risk factors and many study subjects, analysts can define risk strata and examine the effect of the intervention separately within strata, eliminating the need to model the relationship between the risk factors and the outcome. For example, if age and sex are the only two risk factors considered important, one might compare the outcome of persons receiving/not receiving the intervention separately by sex within each ten-year age category. With many risk factors, we quickly have many strata with too few intervention and control subjects for meaningful analysis. A correctly specified multivariable model solves the problem of having too many small strata. However, it is not easy to prove that such models are sufficiently well specified. Propensity score methods protect against model-induced bias in comparisons of outcomes in groups with notably different distributions of baseline characteristics. Classic articles on the propensity score include Rosenbaum and Rubin (1983, 1984, 1985), Rubin and Thomas (1996), Rubin (1997), and D'Agostino (1998).

We first illustrate how model misspecification and different covariate distributions can lead to incorrect estimates of the effect of an intervention. We then discuss how propensity scores can be used to produce more credible findings, discuss the types of variables that should be included in propensity score models, and briefly describe several applications in the published literature.

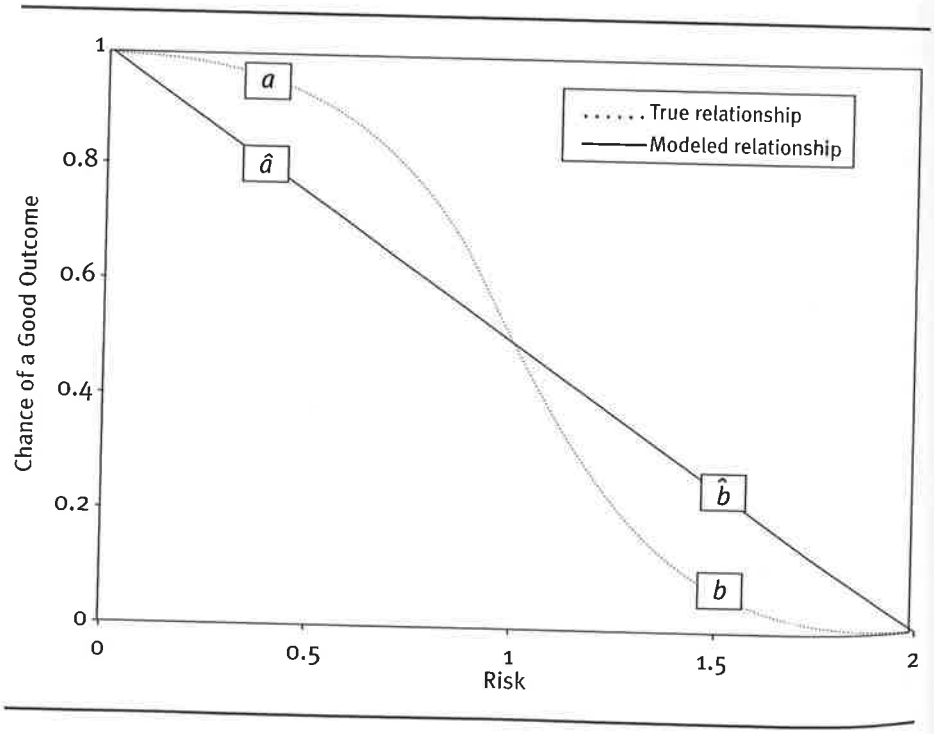
Hypothetical Example

Assume we have a summary risk score that varies from 0 (lowest risk) to 2, an outcome measure that varies from 0 (worst outcome) to 1 (best), and an S-shaped curve (as shown by the dotted line in Exhibit 11.1) representing the unknown but true relationship between risk and outcome. At risk levels below 0.5, outcomes are good and worsen slowly with increasing risk. At risk levels between 0.5 and approximately 1.5, outcomes decline linearly and much more steeply than they do at lower risk levels. At risk levels above 1.5, outcomes are poor but decline gradually as risk increases. The solid straight line in Exhibit 11.1 shows an incorrectly specified (straight-line) model of the relationship between risk and outcomes. At lower risk levels, the model-predicted outcome

is worse (lower) than the true outcome; at high risk levels, the model-predicted outcome is better (higher) than the true outcome.

Assume that persons at high risk for a bad outcome differentially receive an intervention. Specifically, suppose that in the situation depicted in Exhibit 11.1, individuals receiving the intervention mostly have risk scores around 1.5, while risk scores in the control group are mostly around 0.4. Due to baseline risk differences alone, about 95 percent of the controls would have a good outcome (see *a* in Exhibit 11.1), but only 5 percent of those destined to receive the intervention (see *b* in Exhibit 11.1) would have a good outcome. However, although the underlying difference in outcomes due to differential baseline risk is approximately 90 percent, our model (illustrated by the straight line in Exhibit 11.1) underestimates that difference. Its predictions are overly pessimistic about the expected outcome for those at lower risk levels (where most of the control cases are) and overly optimistic about the expected outcome for those at the higher risk levels (where most of the intervention cases are). Specifically, our model expects 80 percent good outcomes for the control group (see $\hat{a}$ in Exhibit 11.1) and 25 percent good outcomes for the intervention group (see $\hat{b}$). Thus, in the case of a null-effect intervention, the model expects good outcomes to be only 55 percentage points less common for the intervened patients than for untreated patients, understating the true baseline disadvantage for the intervention group by 35 percentage points.

EXHIBIT 11.1 True and Modeled Relationships Between Risk and Outcome



Suppose that the intervention is incredibly successful: Those who receive it are 40 percentage points more likely to have a good outcome than they would have been without it. Had the intervention produced no effect, the two success rates would have been 95 percent and 5 percent; due to the intervention, they are 95 percent and 45 percent. Although the intervention works, the raw data associate the intervention with a decrease of 50 percentage points in good outcomes. Using the inadequate model to adjust for risk, we would think that the intervention is helpful but only associated with a 5 percentage point improvement (because the model "expects" a null intervention to have 55 percentage points less success). While this example is extreme, the lesson is that, especially when dealing with a complex, multivariable situation, relying on "modeling heroics" to compare outcomes for very different kinds of people is precarious. When we try to compare groups that are far from comparable, the validity of a model-based finding relies on the hope that the ways in which the model is imperfect affect the intervention and control groups similarly.

Using Propensity Scores

If we could compare the outcomes of subgroups of study and control patients with similar risk profiles, model imperfections would matter less. Propensity scores help identify such subgroups, allowing us to approximate, in an observational setting, some of the virtues of a designed trial. In a prospectively designed study, subjects are often specially selected to establish the best conditions for answering a highly specific study question, such as, "For the kind of person enrolled in this study, will a new intervention lead to a better outcome than usual care?" When eligible subjects are randomly selected to receive either the intervention (treatment) or usual care (control), both treated and control subjects will be similar on all covariates (including those that are not measured). Thus, in a planned study with random assignment, the unadjusted difference in outcomes between intervention and control patients should largely reflect the treatment effect, although standard regression modeling is often used to reduce unexplained variation in the outcome and increase the chance that an effective intervention is identified as statistically significant. Propensity score methods can be used to make the observational data used in a study satisfy the condition that the treatment and intervention groups are at least similar on all **measured** covariates. Several excellent resources explaining the role and facilitating the use of propensity scores in observational studies are available. For example, Ho and colleagues (2007) propose a unified approach that enables researchers to preprocess data using propensity score methods, after which standard parametric modeling techniques can be used to improve precision.

Consider a multivariable logistic regression model that uses risk factors to predict receipt of the intervention—that is, the dependent variable equals

1 if the intervention is received and 0 otherwise. The propensity score for the i^{th} subject is the $PRED_i$ from this regression: the predicted probability that the i^{th} person receives the intervention. If subsets of the intervention and control groups that have similar distributions of the propensity score are chosen, these subsets also will have similar distributions of the risk factors considered by the propensity score model. There are several common ways to identify subsets of study and control group members with similar propensity scores. The simplest uses random sampling within propensity score categories, as follows:

1. Develop a logistic regression model to predict the probability of receiving the intervention as a function of risk factors (later in this chapter we offer guidelines for selecting the specific risk factors used in the propensity score model). As noted, the predicted probability of receiving the intervention is the propensity score.
2. Rank the population from lowest to highest propensity score.
3. Divide the population into groups with similar propensity scores. Here, we assume that propensity score quintiles, which typically work well, are used.
4. Within each propensity score quintile, divide people into an intervention group and a control group (see Exhibit 11.2). Controls are most common in the lower quintiles (where the predicted probability of receiving the treatment is smallest) and least common in the higher quintiles.
5. Sample equal numbers of intervention cases and control subjects from each quintile. As shown in Exhibit 11.2, the number of subjects in the smaller group in each quintile limits the number that can be sampled. For example, in the lowest quintile the intervention group has 12 subjects and the control group has 81, so we select all 12 intervention subjects and sample 12 of the 81 controls, for a total of 24 individuals. In the highest quintile the intervention group has 78 subjects and the control group has 8, so we select all 8 control subjects and 8 of the 78 intervention subjects, for a total of 16. The combined sample of 206 subjects in Exhibit 11.2, then, has equal-sized control and intervention groups that have identical distributions of propensity score quintiles.[9]
6. Examine the distribution of risk factors between intervention and control subjects in the sample to see how similar their measured risk factors are (we discuss this topic in more detail later in this chapter).
7. Fit a new model to predict the effect of the intervention on the combined sample. Because the intervention and control subjects in the sample are similar with respect to baseline factors, raw and risk-adjusted estimates of the treatment's effect should be similar. Though modeling can reduce bias produced by residual differences between the sampled

Quintiles of Propensity Score	Number of Intervention Cases	Number of Control Cases	Analysis Sample
<20%	12	81	24
20–40%	30	67	60
40–60%	44	38	76
60–80%	53	15	30
>80%	78	8	16
Total	217	209	206

EXHIBIT 11.2 Example of Sampling Within Propensity Score Quintiles

groups, the main purpose of modeling at this stage is to produce a more accurate estimate: a narrower confidence interval for the effect of the treatment.

Several points should be noted about this approach. First, it is useful to estimate the effect of the treatment separately within propensity score quintiles if enough data are available. If treatments are allocated rationally, patients with the highest propensity scores should experience the greatest benefits. Also, this estimate could provide insight into how effective a treatment would be for patients who are currently unlikely to receive it. Second, dividing the propensity score into quintiles and then sampling from quintiles is generally adequate to produce balanced samples; in particular, analysts need not create and sample from finer propensity score categories (Rosenbaum and Rubin 1983). Perfectly matched samples are unnecessary; they just need to have reasonably similar distributions of factors that importantly affect the outcome of interest. Finally, propensity score methods focus on how intervention and control subjects in the full population differ on measured factors. These differences may help explain different findings between standard and propensity score analyses and could reveal that the intervention is much more effective for some kinds of patients than for others.

Another simple approach to creating subsets of similar intervention and control subjects is "nearest available matching on the estimated propensity score" (Rosenbaum and Rubin 1985, 35). Intervention group participants are randomly ordered. In the initial step, the first intervention group participant is matched with the control group participant whose propensity score is closest, and the pair is set aside. At each step, the next remaining study group participant is matched with the remaining control group participant whose propensity score is closest and the pair set aside until all study group participants are matched. This approach is an example of the "greedy algorithm"—once the nearest available match has been selected, the pairing

is not reconsidered. The approach can be improved by setting a maximum acceptable distance between intervention and control subjects' propensity scores (called a *caliper*). When no control patient within the caliper is available for matching, the intervention case is discarded. This restriction prevents comparing outcomes between intervention subjects and "extremely dissimilar" control subjects.[10] Note that both the greedy algorithm approach and sampling within quantiles have a random component. As a result, somewhat different samples and (especially when data sets are small) potentially different findings could emerge when two analysts use either method on identical data—a conceptually undesirable property, but one that rarely causes any practical difficulty. However, the problem can be avoided by using more sophisticated matching approaches, such as Mahalanobis metric matching either with or without calipers (Rosenbaum and Rubin 1985) and optimization models (Rosenbaum 1989).

Inverse probability of treatment weighted (IPTW) estimation is an elegant alternative to matching; the weights equalize the "effective contributions" of intervention and control cases across all kinds of patients by down-weighting groups of observations rather than discarding them (Rosenbaum 1987; Joffe et al. 2004). In IPTW estimation, all (or nearly all) observations are used in the multivariable model, but some observations count more than others. First, let ps_i = the estimated propensity score for case i. If case i receives the intervention, the case is weighted by $1/ps_i$; if case i does not receive the intervention, it is weighted by $1/(1 - ps_i)$. Thus, subjects with high propensity scores who receive the intervention (the expected cases) are weighted less than those with high propensity scores who do not receive it (the unexpected cases), and subjects with low propensity scores who do not receive the intervention (an expected situation) are weighted less than those with low propensity scores who receive the intervention (unexpected). The weights must be properly accounted for when one is estimating variances and significance levels (Joffe et al. 2004). An IPTW analysis uses all the data and readily leads to a single estimate of the size of the intervention's effect, but this virtue may mask important features. In contrast, propensity score matching encourages analysts to look at the data and the propensity score distributions and perhaps see that estimated effect sizes differ across groups of patients or that some "oddball" cases (which would be highly influential in an IPTW estimate) might be skewing the effect-size estimate or unduly inflating its variance. Because subjects with extreme weights are highly unusual and may reflect data errors, findings that depend on them are suspect. It is often a good idea to eliminate these highly influential cases from the analysis.

The most common application of the propensity score in the medical literature is to use it as a covariate in addition to other risk factors in a standard multivariable regression on the full study sample (Shah et al. 2005; Stürmer et al. 2006). In such settings, the propensity score is used to

summarize in a single dimension the many ways in which treated and untreated subjects' risk factors differ. While this use works as a "variable reduction technique" for a modest-sized data set including many predictors on which the intervention and control groups differ, it does not protect against model misspecification.[11]

Austin (2009b) compared the ability of the four propensity score methods just discussed to balance measured covariates between study and control subjects using real and simulated data. The real data were from approximately 7,600 patients discharged with a diagnosis of heart failure from 103 acute care hospitals in Ontario. He studied two interventions: receipt at discharge of a prescription for (1) a β-blocker and (2) an angiotensin-converting enzyme inhibitor. For each drug type, systematic differences appeared between treated and untreated subjects on several baseline characteristics. While all four propensity score methods removed a large proportion of the systematic differences between groups, propensity score matching and weighting using IPTW performed better than did stratification and covariate adjustment. Matching and IPTW "eliminated virtually all observed systematic differences" (Austin 2009b, 669).

Real-World Example Using Propensity Scores

We used propensity score matching to compare outcomes for two modalities of substance abuse detoxification ("detox") in the publicly funded substance abuse treatment system in Boston: (1) acupuncture, which included daily sessions during the acute detox phase, followed by sessions two to three times per week for three to six months plus maintenance and motivational counseling, usually in groups, and (2) residential detox, which lasted approximately one week, and encouragement at discharge to seek post-detox treatment. The outcome of key interest was whether clients were readmitted for detox within six months of an index detox admission (Shwartz et al. 1999).

Acupuncture clients differed substantially from those who underwent residential detox; the acupuncture patients typically fared better. For example, 13 percent of acupuncture patients were college graduates compared to 4 percent of residential detox patients, 57 percent versus 13 percent were employed, 15 percent versus 3 percent had private insurance, 76 percent versus 55 percent lived with others, 3 percent versus 30 percent lived in a shelter, and 19 percent versus 43 percent had other detox admissions in the year prior to the index admission. Among 1,104 acupuncture clients and 6,907 residential detox controls, we matched 740 subjects of each type on propensity score by sampling within propensity score deciles, a matching method we might reconsider if we were doing the study today. These two groups of 740 persons had largely comparable risk factors. For example, 7 percent in both groups had college degrees, 42 percent of acupuncture versus 41 percent of residential detox patients were employed, 6 percent versus 9

percent had private insurance, 77 percent versus 72 percent lived with others, 4 percent versus 5 percent lived in a shelter, and 26 percent versus 27 percent had other detox admissions in the last year.

The effectiveness of acupuncture for preventing a detox readmission was relatively similar when the analysis was performed on the full sample (OR 0.71, 95 percent confidence interval 0.53 to 0.95) and on the propensity score–matched subsamples (OR 0.61, 95 percent confidence interval 0.39 to 0.94). This similarity may relate to the extremely high predictive power (c-statistic = 0.96) of the original model, which, fit to the full sample, left little room for model misspecification.

As noted, analysts can use propensity scores in several ways. In the acupuncture project, we used propensity scores to sample patients and then estimated the intervention's effect on the sample. This approach is transparent; it highlights the similarities and differences between the individuals who did and did not receive the intervention. We had started with a standard multivariable analysis comparing the acupuncture and residential treatment program outcomes, but reviewers criticized these analyses, arguing that users of the two treatment modalities were so fundamentally different that their outcomes could not be compared. However, we found 740 residential detox patients whose risk factors were closely similar to those of a subset of 740 acupuncture patients. The similarity between the groups' risk factors reduced concerns that the original finding was an artifact of "heroic modeling" between groups that were too dissimilar to usefully compare. Ultimately, comparing the outcomes of subsets of persons who had very similar measured covariates but were receiving either residential detox or acupuncture was more convincing than relying on risk adjustment to compare the outcomes of full samples of these different populations.

Variables That Should Be Included in Propensity Score Models

Because the dependent variable in a propensity score model is whether an intervention is received, one might think the most important criterion in selecting risk factors for inclusion in the model is the association of the risk factors with receipt of the intervention and that the c-statistic is a useful summary measure of the propensity score model. However, this assumption is not the case (Rubin and Thomas 1996; Rubin 1997). Brookhart and colleagues (2006b) conducted simulations that provide guidance on selection of variables for inclusion in propensity score models. Next we illustrate the effect of including certain types of variables in the models.

Risk factors related to receipt of both the intervention and the outcome are called confounders. To illustrate the effect of confounders and adjustment for confounders, we find it useful to take a step back from propensity scores and consider direct matching of risk factors to balance risk factors between intervention and nonintervention groups. Consider a (harmful)

risk factor, which we will call RF1. This risk factor can take one of two values: present or not. Exhibit 11.3A shows that the distribution of RF1 differs by intervention status, and Exhibit 11.3B shows the percentage of people experiencing a good outcome by risk factor and intervention status.

RF1 is a confounder; it is distributed differently among those receiving the intervention (1/3 have RF1) and those not receiving it (2/3 have it), and in both groups, those who have it do worse than those who do not. Note that overall, 70 percent of those receiving the intervention have a good outcome versus only 40 percent of those not receiving the intervention. We can adjust for the confounder by matching—that is, by selecting equal numbers in the high- and low-risk groups from both the intervention and nonintervention groups (see Exhibit 11.3C). In the matched sample, 65 percent ([800 + 500]/2,000) of those in the intervention group and 45 percent ([300 + 600]/2,000) of those not in the intervention group have a good outcome; the intervention now appears to be effective, but less so than it appeared to be before adjusting for the confounder.

What happens when we match on a second risk factor, RF2, which is related to whether a person receives the intervention but not to the outcome? Specifically, among those who receive the intervention, 80 percent have RF2, and among those who do not receive the intervention, 20 percent have it. In addition, let us assume that RF1 and RF2 are independent. Exhibit 11.3D shows the numbers categorized by each of the risk factors and intervention status. Exhibit 11.3E shows sample numbers (and the numbers experiencing a good outcome) by intervention status, after matching on both RF1 and RF2. Overall, 65 percent (780/1,200) of those receiving the intervention and 45 percent (540/1,200) of those not receiving it had a good outcome, exactly the same percentages we arrived at when we matched on RF1 alone. However, now the estimates are based on only 1,200 in each group rather than 2,000. This smaller sample size makes the uncertainty associated with each estimate greater, increasing the chance of falsely accepting the null hypothesis.

The same happens when a variable related to the likelihood of receiving the intervention, but not (or only minimally) related to the outcome, is added to a propensity score model. It reduces the sample sizes that can be matched (widening confidence intervals) without removing bias from the point estimate of the effect of the intervention. Hence, while such variables cause little harm in a regression model (only "costing" one degree of freedom), they can seriously reduce the power of propensity score analyses. Variables uncorrelated with the outcome should not be included in propensity score models.

Distribution of Covariates in Propensity Score–Matched Groups

As noted, the principal goal of propensity score matching is to achieve similarity in the distribution of baseline risk factors between those who receive an

EXHIBIT 11.3A
Number of People Categorized by RF1 and Intervention Status

RF1	Intervention	No Intervention
Present	1,000	2,000
Absent	2,000	1,000
All	3,000	3,000

EXHIBIT 11.3B
Percentage (Number) of People Experiencing a Good Outcome by RF1 and Intervention Status

RF1	Intervention	No Intervention
Present	50% (500)	30% (600)
Absent	80% (1,600)	60% (600)
All	70% (2,100)	40% (1,200)

EXHIBIT 11.3C
Number of People Categorized by RF1 and Intervention Status, After Matching Cases

RF1	Intervention	No Intervention
Present	1,000	1,000
Absent	1,000	1,000

EXHIBIT 11.3D
Number of People Categorized by RFs 1 and 2 and by Whether They Receive an Intervention

RF1	RF2	Intervention	No Intervention
Present	Present	800	400
Present	Absent	200	1,600
Absent	Present	1,600	200
Absent	Absent	400	800

EXHIBIT 11.3E
Number of People (and Number Experiencing a Good Outcome) by RFs 1 and 2 and Intervention Status, After Matching on Both Risk Factors

RF1	RF2	Intervention	No Intervention
Present	Present	400 (200)	400 (120)
Present	Absent	200 (100)	200 (60)
Absent	Present	200 (160)	200 (120)
Absent	Absent	400 (320)	400 (240)

intervention and those who do not. Austin (2009a) discusses several ways to examine the extent to which this similarity is achieved. We review some of these methods in the following paragraphs.

First, it is valuable to show mean values (or prevalence rates) of important predictor variables (*X*s) in (1) the full population of those who received and those who did not receive the intervention and (2) the propensity-matched samples of these two groups. Such data provide general reassurance that large covariate imbalances in the original populations have been reduced to small differences in the matched samples used in analyses in which the goal is to isolate the effect of treatment specifically for the kinds of people who were treated.

The decision that the covariate means in two groups are similar enough is clearly subjective. However, examination of standardized differences clarifies which variables originally differed the most, facilitates identifying the variables that exhibit the greatest imbalances even after matching, and helps one determine whether further attempts to achieve balance should be made. Standardized differences enable plausible comparisons of variables that would otherwise be minimally comparable, such as years of age, dollars of health care spending, and the presence of specified comorbidities, for the specific purpose of determining the variables on which the matched treated and control subjects differ the most. For each variable X, we are interested in the observed difference of the mean of X between the two groups: $Diff(X) = \overline{X}_1 - \overline{X}_2$. Clearly, $Diff(X)$ is expressed in the units in which X is measured. However, when $Diff(X)$ is divided by a measure of how spread out the *X*s are (e.g., by an estimated standard deviation, or $SD(X)$ = the square root of the variance of X), it becomes "unit-less" and therefore comparable to other standardized differences. Because X may vary differently within the two groups, Austin (2009a) standardizes the difference by dividing $Diff(X)$ by the square root of the average of the two observed variances:

$$Standardized\ Diff(X) = (\overline{X}_1 - \overline{X}_2) \div \sqrt{\frac{s_1^2 + s_2^2}{2}},$$

where the s_i^2s are sample variances. For a dichotomous risk factor, this equation looks like:

$$Standardized\ Diff = (p_1 - p_2) \div \sqrt{\frac{p_1(1 - p_1) + p_2(1 - p_2)}{2}},$$

where p_1 and p_2 are the observed prevalences in the two groups.

Standardized differences are often called effect sizes. Cohen (1988) introduced the widely adopted terms *small-*, *medium-*, and *large-effect sizes* to describe standardized differences of 0.2, 0.5, and 0.8, respectively. Standardized

differences can be portrayed, as in Figure 1 of the article by Austin (2009a, 3089), to illustrate the improvement in risk factor balance between an intervention group and a control group after propensity score matching.

Two groups can have similar means but otherwise different distributions of a covariate. Ideally, the ratio of the variances in the groups, s_1^2/s_0^2, should be close to 1.0 after propensity score matching. Side-by-side box plots of risk factors in the two groups can also be examined to more fully compare similarity of distributions and demonstrate improved similarity after propensity score matching. Ho and colleagues (2007) suggest using a Q-Q plot in which the quantiles of each distribution are plotted against each other. Many statistical software packages generate density plots for continuous variables (e.g., SAS, Stata, R).[12] By laying the density plots of a risk factor from two groups over each other, we can easily see when and how their full distributions differ.

Final Thoughts About Propensity Scores

The use of propensity scores has grown rapidly and is now a standard part of multivariable modeling used to assess treatment effectiveness, particularly in studies related to heart disease. Reasons for this growing use were nicely summarized by Rubin (1997, 763) in an early, nontechnical article:

> [Standard multivariable models are] fraught with pitfalls because of their reliance on unwarranted assumptions and extrapolations without any warning. . . . One critical advantage of propensity score methods is that they can warn the investigator that, because of inadequately overlapping covariate distributions, a particular database cannot address the causal question at hand without relying on untrustworthy model-dependent extrapolation. . . . Because of this advantage, any causal questions put to a large database should be first approached using propensity score methods to see whether the question can be legitimately addressed. If so, subclassification on a well-estimated propensity score can be used to provide reliable results. . . . After that, modeling can play a useful role. For example, standard statistical models . . . can be safely applied within propensity score subclasses to adjust for minor within-subclass differences.

But do propensity score methods provide meaningfully different conclusions from those reached using conventional multivariable models? Stürmer and colleagues (2006, 437) asked this question, concluding: "Publication of results based on propensity score methods has increased dramatically, but there is little evidence that these methods yield substantially different estimates compared with conventional multivariable methods." The June 2, 2010 issue of the *Journal of the American Medical Association* supports this conclusion. The authors of all three articles evaluating heart disease–related interventions (i.e., Lambert et al. 2010; Marso et al. 2010; Kilgannon et al. 2010) used propensity scores, one in the main analysis and two in sensitivity analyses. Lambert and

colleagues (2010, 2153), who used propensity scores for sensitivity analysis, noted, "The resultant ORs from the matched propensity analysis . . . were not substantially different than those yielded by the multivariate logistic regression models." Kilgannon and colleagues (2010, 2169), who also used propensity scores for sensitivity analysis, reported, "In the sensitivity analysis adjusting the model for propensity scores, the OR and 95% CIs . . . did not change." In contrast, Maciejewski and colleagues (2011) did find a difference when they used propensity score methods. They used a proportional hazards model to examine survival following bariatric surgery among high-risk patients treated at Veterans Affairs medical centers. Among the 850 veterans who had bariatric surgery, compared to more than 41,000 controls, bariatric surgery was associated with reduced mortality (hazard rate [HR] = 0.80, 95 percent confidence interval [CI] 0.63–0.995). (Note the need to report the high end of the confidence interval to three decimal places to demonstrate statistical significance.) In an analysis of 1,694 propensity score–matched patients, the point estimate for the effect of bariatric surgery was closer to 1.00 (indicating no effect) and had a substantially wider confidence interval, thereby erasing the perception of any effect (HR = 0.94, CI 0.64–1.39). These results suggest that the full set of control cases used in the traditional analysis were notably less healthy than those who received bariatric surgery, producing—in that analysis—an overly optimistic estimate of the effect of the surgery.

Regardless of whether propensity score methods change overall conclusions, they provide an important opportunity to explore how well the distributions of risk factors overlap for those receiving and not receiving an intervention. These methods also remind us that for the types of patients who either never or always receive the intervention, the data contain essentially no information relating to the intervention's effectiveness. If patients for whom there is overlap are a much restricted subset of the total population, even when a traditional analysis yields "the same answer" yielded by an analysis based on a propensity score–matched sample of patients, only the latter clearly exposes the reality that the findings apply to patients only where there is overlap. For this reason, propensity score analyses are generally more credible than traditional analyses.

Instrumental Variables

Standard multivariable modeling and propensity score methods are powerful tools for distinguishing the effect of an intervention from the effect of differences in baseline risk. However, these techniques can still mislead in the presence of **unmeasured** risk factors that both (1) strongly affect outcomes and (2) are unevenly distributed between intervention and control subjects. Instrumental variables, used extensively in econometrics (e.g., Greene 2003), can assess intervention effects in this seemingly hopeless situation.

We can account for the effect of an unmeasured risk factor with a suitable instrument for studying the intervention. An *instrument* is an observed variable that meets two criteria: (1) The variable is associated with the likelihood of receiving the intervention, and (2) the variable is not directly associated with the likelihood of a good outcome. Angrist, Imbens, and Rubin (1996) state more formally the assumptions required for a variable to be a valid instrument.

In the next section, we offer a simplified hypothetical example (based on Greenland 2000a) of the instrumental variable (IV) approach in the context of an RCT in which compliance with the intervention was incomplete. The arm of the trial to which the person is randomized is an excellent instrument. Persons randomized to the study group are much more likely to receive the intervention than are those randomized to the control group. However, the instrument is not directly associated with the likelihood of a good outcome because those randomized to the study group are similar to those randomized to the control group in terms of both measured and unmeasured risk factors. Therefore, in the absence of the intervention, their outcomes should be similar. Later, we discuss several examples from the health services research literature in which IVs were used to assess the effectiveness of an intervention.

Hypothetical Example

Consider 2,000 people, one half randomized to a study group that receives an intervention and the other half to a control group. Of the 1,000 people in the study group, only 600 comply with the intervention. None of those in the control group receives the intervention. The top table in Exhibit 11.4 shows the number of people and number of deaths in each group.

A standard analysis of an RCT called an *intention-to-treat analysis* compares outcomes among all subjects in the study group, whether they received the intervention or not, to outcomes among all those in the control group. In our example, the death rate is 5.2 percent in the study group (52/1,000) and 8.8 percent in the control group. An intention-to-treat analysis finds a 41 percent reduction in deaths due to the intervention.

This analysis suggests that offering the intervention led to 41 percent fewer deaths than would have occurred had the intervention not been offered. It does not indicate what the effect of the intervention would have been if everyone in the study group had complied with the intervention. As rates of noncompliance rise, intention-to-treat analyses increasingly underestimate the true effect of the intervention **on those who receive it**.

One alternative to an intention-to-treat analysis is an *as-treated analysis*, which compares the outcomes of cases who received the intervention (the compliers) to those in the control group who did not. In our example, 12 of the 600 compliers died—a 2 percent death rate. This rate is a 77 percent reduction from the 8.8 percent death rate in the control group. The problem

EXHIBIT 11.4 Randomized Controlled Trial (RCT) Example Illustrating Instrumental Variables

What we observe from the RCT

	Study Group		Control Group	
	Compliers	**Noncompliers**	**All**	**All**
No. of patients	600	400	1,000	1,000
No. of deaths	12	40	52	88
Death rate	2%	10%	5.2%	8.8%

Plausible underlying reality*

	Study Group		Control Group	
	Compliers	**Noncompliers**	**Compliers**	**Noncompliers**
No. of patients	600	400	600	400
No. of deaths	12	40	??	40
Death rate	2%	10%	?%	10%

*The total number of deaths in the control group = 88, so it is reasonable to impute 48 for ??, which makes ?% equal 8%.

Data from the RCT

	Study Group	Control Group
No. of patients	1,000	1,000
No. of deaths	52	88
No. treated	600	0

with as-treated analyses is that compliers are often a nonrandom subset of the study group. They are usually better educated, have higher incomes, are healthier, and are expected to do better even without the intervention. As-treated analyses generally overstate treatment benefits because of this "healthy volunteer" bias. Because we do not know to what extent the study group compliers differ from the control group in terms of both measured and unmeasured risk factors, such studies no longer carry the credibility conferred by the original randomization. For this reason, as-treated analyses, once commonly performed, have fallen into disfavor.

A third approach builds on a "counterfactual" or potential-outcomes framework to estimate causal effects from observational data (Rubin 1974,

1977, 1978). The key idea is that individuals in a particular state (in our example, a control group or an intervention group) have potential outcomes in other states (in this case, the state to which they were not randomized). In our example, the control group's counterfactual state is assignment to the study group. What are reasonable potential outcomes for the control group in its counterfactual state? Randomization gives us confidence that the control group is similar to the study group. Hence, it seems reasonable to assume that had the control group been offered the intervention, 600 people would have complied with the intervention and 400 would not have. Furthermore, it is reasonable to assume that the 400 noncompliers in the control group would be similar to the 400 noncompliers in the study group. Hence, we would expect 40 deaths among these control group noncompliers in their counterfactual state (see the second table in Exhibit 11.4). Because there were a total of 88 deaths in the control group, we would expect the rest of the deaths (48) to occur among the 600 counterfactual control group compliers. Control group counterfactual compliers can be legitimately compared to study group actual compliers. Among study group compliers, 12 deaths occurred, which suggests that the intervention reduced deaths by 75 percent in this group of 600 treated individuals.

The logic of IV analysis is similar to the analysis in the last paragraph, although the steps are different. The third table in Exhibit 11.4 shows the data available from the RCT. The instrumental variable is group assignment. As the value of the IV switches from control to study group, two things happen:

1. The number of deaths declines from 88 to 52.
2. The number of persons receiving the intervention increases from 0 to 600.

What could account for the decline in the number of deaths? It is not caused by differences in either measured or unmeasured risk factors because randomization makes them similar for each level of the instrumental variable. The lower death rate must be due to the increased number receiving the intervention in the study group.

The estimate of treatment effectiveness in an IV analysis is

$$\frac{\textit{Change in likelihood of dying, by level of the IV}}{\textit{Change in likelihood of receiving treatment, by level of the IV}}.$$

In our example, $(0.052 - 0.088)/(0.60 - 0.00) = -0.06$.

To interpret the −0.06, note that the study saves 36 lives (88 − 52) by providing the intervention to 600 compliers (i.e., for every complier, the intervention saves 0.06 lives [36/600]). We know from the RCT results that 12 deaths occurred among the 600 compliers. Because 36 lives were saved, we would have expected 12 + 36 = 48 deaths among the 600 compliers if they had

not been given the intervention. The intervention reduced the number of deaths per 600 from 48 to 12 (a 75 percent reduction), which is consistent with the effectiveness estimate using the counterfactual logic presented earlier.

The IV estimate of intervention effect does not apply to all persons but only to the types of patients who would shift from not having the intervention to having the intervention because of their IV level. In our example, these types are the people similar to the 600 compliers who had been given the intervention. Whereas an intention-to-treat analysis estimates the effect of offering an intervention to an entire population, an IV analysis estimates the effect of that intervention on "marginal patients"—those who would receive the treatment at one level of the IV but not at another.

The IV estimate of effectiveness just illustrated can also be calculated using a two-stage model. Unlike the simple calculation presented earlier, the two-stage approach generalizes to more complex situations. In stage 1, we model the probability of the intervention as a linear function of the IV variable, Z:

$$PRED(I) = a + bZ,$$

where $I = 1$ if the person is in the intervention group and 0 otherwise. In this example, $PRED(I) = 0.6$ when $Z = 1$ and 0 otherwise. Hence, $a = 0$ (the probability of receiving the intervention for the person in the control group) and $b = 0.6$ (the increased probability of receiving the intervention for a person in the study group). In a more complex situation, $PRED(I)$ would be modeled as a function of measured risk factors in addition to the IV variable.

In stage 2, we use a simple linear model to predict the outcome of interest, in this case the probability of death, as follows:

$$PRED(Death) = c + [d \times PRED(I)].$$

The coefficient d is the effect on $PRED(Death)$ as $PRED(I)$ changes from 0 to 1. In our example, $c = 0.088$ and $d = -0.06$, just as in our earlier calculation.[13] This two-stage approach enables important measured risk factors to be added as covariates at each stage.

We used a linear model for the dichotomous outcome *Death* because it most clearly illustrates the connection between the model's coefficients and estimates in the preceding numerical example. Analysts often use a linear model in the same situation because standard statistical software packages readily accommodate two-stage linear modeling but not nonlinear modeling. While a linear model is theoretically less satisfactory than, say, a logistic model for predicting death, in practice it works fairly well. When using a nonlinear model at the second stage, Terza, Basu, and Rathouz (2008) found that using both I and the residuals from the stage 1 model as predictors at the second stage rather than using $PRED(I)$ produced a less biased estimate of the effectiveness of I.

The assumption underlying IV analysis is that groups defined by different IV states are comparable in terms of unmeasured risk factors. In our example, the IV assumption enabled us to estimate that 1,000 control subjects would have produced 400 noncompliers and 40 deaths among these noncompliers had the control group members been given the treatment. If groups in different IV states are not comparable, results from an IV analysis may be misleading. Unfortunately, there is no direct test of this critical IV assumption. Next, in describing applications in the literature, we consider how respective authors have examined the validity of the IV assumption.

Examples of IV Analyses

McClellan, McNeil, and Newhouse (1994) published a classic study describing the use of IVs in health services research. They examined the value of intensive treatment of acute myocardial infarction (AMI) for reducing mortality among elderly patients. The standard approach was to compare risk-adjusted mortality of patients who did and did not receive intensive treatment (i.e., cardiac catheterization, coronary revascularization, and/or other major surgical procedures). However, these two groups had large differences in baseline risk, suggesting that other important unmeasured differences might be present. For example, 2.8 percent of the patients who received catheterization within 90 days had cerebrovascular disease versus 5.4 percent among those who did not. One can easily imagine how these differences might arise. Consider two people, one relatively hardy and the other frail, with the same measured risk factors. If only one is treated aggressively, it is likely to be the stronger individual, who probably would fare better than the frailer patient, even without treatment. Suspecting that such unmeasured risk factors had produced inappropriately optimistic estimates of treatment effects, McClellan and colleagues conducted an IV analysis.

Their instrument was the extra distance a patient must travel to the nearest hospital that offers intensive treatment (beyond the distance to the nearest hospital). They selected this variable as their instrument on the basis of the following assumptions:

1. The shorter the extra distance, the more likely the patient is to receive intensive treatment.
2. Patients for whom the extra distance is short have unmeasured risk factors similar to those of patients for whom the extra distance is long.

The data supported the first assumption. For example, 34 percent of patients whose extra distance was less than 2.5 miles received catheterization within seven days versus 5 percent of those requiring longer travel. However, the second assumption—the key to the validity of an IV estimate—could not be directly verified. Examining the plausibility of the IV assumption commonly

involves comparing the distributions of measured characteristics in the IV-defined groups, which were fairly similar in this study. For example, in both groups, 4.8 percent of the patients had cerebrovascular disease. If study groups have different patterns of measured characteristics, their unmeasured characteristics are unlikely to be similar. Comparable distributions of measured risk factors across IV states make the IV assumption more plausible.

A standard risk-adjusted analysis estimated a 28 percent reduction in four-year mortality among patients who received catheterization. The IV analysis estimated a 6 percent reduction. The smaller estimated effect suggests that catheterized patients were systematically healthier in ways not captured by the measured risk factors. However, as stated earlier, IV estimates pertain only to marginal patients (i.e., those who receive catheterization only when it is more readily available) (Harris and Remler 1998). Among AMI patients for whom place of residence does not affect treatment, the IV analysis says nothing about treatment effects. But finding 34 versus 5 percent seven-day catheterization rates for patients with different IV states tells us that, for many patients, home location strongly influences the chance of receiving early intensive treatment.

In contrast, estimates of effectiveness based on propensity score matching apply to intervention and control subjects who are similar on measured covariates. Thus, estimates of effects from IV and propensity score analyses could differ without either being incorrect. In particular, a plausible explanation exists for at least some of the 28 percent standard effect estimate versus the 6 percent IV estimate. Early aggressive surgical intervention is highly effective for a targeted set of people whose characteristics indicate they almost always receive these therapies. In contrast, the benefit is minimal for people with limited disease, for whom the need for aggressive therapy is less widely accepted and for whom receipt of it depends on its accessibility.

The December 2000 (Part II) issue of *Health Services Research* is devoted to applications of instrumental variables (McClellan and Newhouse 2000). One study (Malkin, Broder, and Keeler 2000) in this issue used a particularly interesting instrument—time—to examine the effect of postpartum length of stay on newborn readmission rates. A standard analysis would compare risk-adjusted readmission rates of short-stay and long-stay newborns. An IV analysis could account for the possibility that these two groups differ in ways not captured by the measured risk factors. One of the instruments used in their analysis was hour of birth: Average length of stay was longer for babies born in the a.m. than for babies born in the p.m. Thus, the IV related directly to the likelihood of the intervention, in this case a long hospital stay. In terms of observable risk factors, babies born in the a.m. and p.m. were fairly similar. This similarity increases our confidence in the key IV assumption that unobservable risk factors do not depend on the value of the instrument. The standard analysis suggested that increasing hospital stays by

12 hours reduces the newborn readmission rate by 0.3 percent. The IV analysis using time of birth as the instrument suggests that increasing length of stay by 12 hours reduces the readmission rate by 0.6 percent, double the rate calculated in the standard analysis. These results suggest that the standard approach underestimates the benefit of a longer stays because babies with longer stays are sicker in ways not fully captured by measured risk factors.

Bao, Duan, and Fox (2006) analyzed the effect of provider advice on smoking cessation. Both smoking cessation and receipt of advice to quit smoking during a medical visit in the last 12 months were determined from the National Health Interview Survey. The investigators were concerned that heavy smokers and those diagnosed with a smoking-related condition may have been more likely than light smokers to receive advice to quit. The first group (heavy smokers) might have a harder time quitting than light smokers, which would cause the analysis to underestimate the effect of the advice; on the other hand, if the first group (which included those diagnosed with a smoking-related condition) was more motivated to quit for health reasons, the analysis might overestimate the effect of the advice. To address potential selection bias in receipt of treatment (i.e., smoking advice), patients were classified as "advised" on the basis of whether they reported receiving advice *on diet and physical activities*—that is, having a clinician who gave different lifestyle advice was used as an IV. It is a good IV to the extent that (1) clinicians who give advice on diet and physical activity are likely to give smoking advice to all smokers and that (2) advice on diet and physical activity does not directly affect smoking cessation. A naïve (standard) analysis provided a highly counterintuitive finding, suggesting 7 percent quit rates among those who received smoking cessation advice and 16 percent quit rates among those who were not given this advice ($p = 0.006$). However, the IV analysis reached the (far more credible) opposite conclusion: Provider advice increases patients' chance of quitting from about 7 percent to 15 percent ($p < 0.001$). The authors conducted several additional analyses to evaluate the validity of their IV. These analyses indicated it was unlikely that more health-minded patients see physicians who practice more preventive care or that differences in the severity of existing conditions accounted for a significant part of the observed effect. They also conducted sensitivity analyses (see discussion later in this chapter), concluding that it was "extremely unlikely that unobserved health conditions and other shocks would have a relative risk" sufficient to be responsible for the estimated impact of cessation advice (Bao, Duan, and Fox 2006, 2128).

Despite the importance of monitoring the safety and effectiveness of prescription medications in widespread use, selection bias and confounding are serious problems that limit the use of observational data for such studies. Brookhart and colleagues (2006a) evaluated the use of physician-specific prescribing preference as an IV in an observational study comparing the

effect of COX-2 inhibitors to nonselective, nonsteroidal anti-inflammatory drugs (NSAIDs) on gastrointestinal (GI) complications. They estimated a physician's prescribing preference (COX-2 inhibitor versus nonselective NSAID) by using the most recent new NSAID prescription written by that physician. If the last new NSAID prescription the physician had written was for a COX-2 inhibitor, the physician was classified as a "COX-2 prescriber" under the premise that he would be more likely to prescribe the same for the next patient. Otherwise, the physician was classified as a "nonselective NSAID prescriber." The important assumption was that the outcome for the next patient was unrelated to the medication prescribed for the last patient—that is, the likelihood that the next patient would develop GI complications was not related to the medication prescribed for the last patient.

The traditional multivariable model analysis found no protective effect from COX-2 in the 120 days following start of the medication. In contrast, the IV analysis attributed a protective effect to COX-2 exposure, similar to what was found in randomized trials. The researchers concluded that their IV method "appears to have substantially reduced bias due to unobserved confounding" (Brookhart et al. 2006a, 268). However, they also cautioned that "physicians who are frequent users of COX-2 inhibitors are seeing higher risk patients" and thus could expect more GI complications (as validated by an association between the IV and some of the observed risk factors). In addition, they concluded that physicians who are heavy prescribers of COX-2 inhibitors may be more likely to prescribe other medications for additional GI protection, and it may be the other medication, not the COX-2 inhibitors, that reduced GI complications. In another study, Rassen and colleagues (2009) also found physician prescribing preference to be a useful IV.

Final Thoughts About IVs

Health services researchers increasingly use IV analyses. Although much of the IV literature is highly technical, the paper by McClellan, McNeil, and Newhouse (1994) is accessible, and we recommend it to researchers new to IV analysis. Other valuable descriptions of this technique include Newhouse and McClellan (1998) and Greenland (2000a). Several articles in *Health Services and Outcomes Research Methodology* (2001, volume 2, issue 3–4) discuss conceptual and practical issues associated with various analytic methods and compare findings from standard, propensity score, and IV methods applied to the same data.

The challenge in an IV analysis is to identify a good instrument. Few instruments can match the benefits of randomized assignment, but plausible instruments can yield useful insight into the effects of interventions. Nevertheless, concerns remain about the inability to evaluate empirically the extent to which a particular variable satisfies the requirements of an instrument. Murray (2006, 130) notes in his article "Avoiding Invalid Instruments and

Coping with Weak Instruments" that "applying instrumental variables persuasively requires imagination, diligence, and sophistication."

IV assumptions are typically justified by examining distributions of measured covariates and discussing the magnitude and direction of plausible differences in unmeasured covariates. To the extent that an IV finding relies on untested assumptions, it is not definitive. However, a sizable difference between the effect estimates from a well-conducted IV analysis and from a standard analysis raises important questions. The search for and testing of hypotheses to explain such differences is one way observational data can elucidate the true object of interest—the intervention's effect on outcomes for particular kinds of patients.

Use of Sensitivity Analysis to Address Unmeasured Confounding

Findings from observational studies can be misleading if important confounders have not been measured. Sensitivity analysis, which broadly refers to examining the data in different ways to see how the study's principal finding changes, may be used to address this threat. For example, when a good instrument is available, an IV analysis provides an alternate view of the evidence. But credible IVs are hard to find, while straightforward sensitivity analysis can (and should) always be performed. It involves considering a range of plausible parameters describing two relationships: (1) the difference in prevalence of an unobserved confounder in the intervention and nonintervention populations and (2) the strength of the effect of the unobserved confounder on the outcome. The idea is to quantify how strong these effects would have to be to substantially change the conclusion suggested by the original analysis. This quantification facilitates judgments about the likely existence of a confounder with the required characteristics.

This approach was first used by Cornfield and colleagues (1959) in their study of the link between smoking and lung cancer. As quoted by Rosenbaum (2010, 105), Cornfield and colleagues wrote:

> If an agent, A, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent B, shows an apparent risk, r, for those exposed to A, relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r. Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is the causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers must be at least 9 times greater than that of nonsmokers.

Cornfield and colleagues concluded that such a large difference in the distribution of a variable like hormone X between smokers and nonsmokers is unlikely, thereby increasing confidence that smoking causes lung cancer.

Several authors have extended and generalized these ideas. Lin, Psaty, and Kronmal (1998) developed a particularly simple set of equations for conducting sensitivity analyses for binary and survival outcomes. Here, we discuss the models for predicting binary outcomes, first illustrating their approach with a binary confounder in a log-linear regression. Consider a dichotomous outcome Y that equals 1 if the event of interest occurs and 0 otherwise; a set of observed covariates X_j, $j = 1, \ldots, J$; a variable I that equals 1 if the subject receives the intervention and 0 otherwise; an unmeasured variable U that equals 1 when the confounder is present and 0 otherwise; and the following model:

$$Prob(Y = 1) = e^{\alpha + \beta I + \gamma_i U + \Sigma_j \theta_j X_j}.$$

Note that $\gamma_i = \gamma_0$ for those who do not receive the intervention and $\gamma_i = \gamma_1$ for those who do. This model is not the more common logistic model in which the left-hand side of the equation is the odds of the outcome rather than its probability. Here, $e^{\gamma_0} = \Gamma_0$ is the effect of the unmeasured confounder on the outcome for the group that does not receive the intervention, $e^{\gamma_1} = \Gamma_1$ is the effect of the confounder on the outcome for the group that receives the intervention, and e^{β} is the intervention's effect. Often, a simplifying assumption is made that these two effects are the same; in this case, their common value is called simply Γ. Because U is unmeasured, Γ cannot be estimated; only the following reduced model can be fit:

$$Prob(Y = 1) = e^{\alpha + \beta I + \Sigma_j \theta_j X_j}.$$

Lin, Psaty, and Kronmal (1998) show that if U is independent of the X_js, both among intervention cases and nonintervention cases, and if β^* captures the apparent effect of the intervention in the observable data and β is its pure, unconfounded effect, then there is a simple relationship between them: $e^{\beta} = e^{\beta^*}/A$. Here, $A = [\Gamma_1 P_1 + (1 - P_1)]/[\Gamma_0 P_0 + (1 - P_0)]$, with P_0 and P_1 being the prevalence of U in the control and intervention groups, respectively.[14] When U confers risk (i.e., when Γ_1 is larger than Γ_0) and U is more prevalent among those with the intervention (i.e., when P_1 is larger than P_0), A is greater than 1, meaning that the intervention's true effect, e^{β}, is less than the observed effect, e^{β^*}. Also, when U confers great risk (i.e., when Γ_1 is much larger than Γ_0), the divisor A approaches its maximum value of P_1/P_0, confirming Cornfield's observation that the amount of apparent risk that can be induced by an unmeasured confounder is bounded by its differential prevalence in the two groups. In the previous formulation, the variances of the estimates of β and β^* are the same, simplifying calculations for point and confidence interval estimates of β for various choices of Γ (or Γ_0 and Γ_1), P_0, and P_1.

In the same article, Lin, Psaty, and Kronmal (1998) use simulation to show that these formulas are good approximations when a logistic regression model is used with a binary U (instead of the previous model), even if the event of interest is not rare. They also derive sensitivity calculations for situations in which the unmeasured confounder is normally distributed and show that these calculations work well for a normally distributed confounder if the event rate is less than 10 percent or if the effect of the unmeasured confounder on the outcome is modest (between 0.5 and 2).

When the confounder is binary (i.e., $U = 1$ if the confounder is present and 0 otherwise), Lin, Psaty, and Kronmal (1998) also consider an alternative formulation in which the effect of U is on I rather than on the outcome. Let ϕ = probability of the intervention; P = overall prevalence of the unmeasured confounder; and ψ, a risk ratio, be $\frac{P(I=1|U=1)}{P(I=1|U=0)}$. From this formulation, the following relationships can be derived:

$$P_1 = \frac{\psi P}{1 - P + \psi P} \quad and \quad P_0 = \frac{P - \phi P_1}{1 - \phi}.$$

Thus, starting with values of ϕ, ψ, and P, we can solve for P_1 and P_0 and enter them into the equation for A (presented earlier) to conduct the sensitivity analyses. Instead of using the risk ratio ψ to express the relationship between U and X, one could also use the odds ratio

$$Q = \frac{\left(\frac{P_1}{1 - P_1}\right)}{\left(\frac{P_0}{1 - P_0}\right)}$$

to express A in terms of Γ_0, Γ_1, Q, P, and ϕ. These formulas provide a straightforward way to conduct sensitivity analyses as described in Rosenbaum and Rubin (1983).

Nichol and colleagues (2007) used the risk ratio formulation to conduct sensitivity analyses in a study of the effectiveness of the influenza vaccine in the community-dwelling elderly population. The main analysis found that vaccination was associated with a 27 percent reduction in the risk of hospitalization for pneumonia or influenza and a 48 percent reduction in the risk of death. In a sensitivity analysis, they considered a binary confounder for which persons with the confounder would be half as likely to be vaccinated and two to three times as likely to be hospitalized or die as persons without the confounder. They varied the prevalence of the confounder from 20 to 60 percent. Under their most pessimistic scenario—that is, triple the risk of the outcome for those with the confounder and 60 percent prevalence of the

confounder—vaccination was associated with a 7 percent reduction in risk of hospitalization for pneumonia or influenza and a 33 percent reduction in the risk of death.

This discussion has focused on analyses that examine the sensitivity of a finding to one type of error: unmeasured confounding. In fact, observational studies have many potential sources of hidden bias. Sensitivity analyses indicate how plausible findings of observational studies are by quantifying how strong a hidden bias must be to substantially change that finding. Several sophisticated algorithms for conducting such analyses are now available in software form, such as Stata's episens program and a SAS macro described by Fox, Lash, and Greenland (2005).

Conclusions

Evidence from observational studies can be—and has been—misinterpreted, leading some to question their value (Sackett et al. 1997). However, these studies often produce useful, credible findings. On the basis of 136 reports from 19 treatments, Benson and Hartz (2000, 1878) "found little evidence that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or qualitatively different from those obtained in randomized, controlled trials." Likewise, after reviewing 99 reports in five clinical areas, Concato, Shah, and Horwitz (2000, 1887) concluded, "The results of well-designed observational studies (with either a cohort or a case-control design) do not systematically overestimate the magnitude of the effects of treatment as compared with those in randomized, controlled trials on the same topic." Ioannidis and colleagues (2001) examined 45 topics in which both randomized controlled trials (n = 240) and nonrandomized studies (n = 168) had been performed. Nonrandomized studies usually showed larger treatment effects. Nonetheless, there were "few differences beyond chance when randomized trials were compared with prospective nonrandomized studies . . . [In addition,] significant between-study variability was seen as frequently across the randomized trials as between the randomized and nonrandomized studies" (Ioannidis et al. 2001, 828).

In this context, observational and RCT studies of hormone replacement therapy (HRT) provided an important cautionary tale from which lessons are still emerging. Analyses of several large observational studies had led to the widespread belief that HRT decreases the risk of coronary heart disease (Stampfer and Colditz 1991; Grady et al. 1992). A decade later, the Women's Health Initiative (WHI), an RCT, found increased coronary heart disease, breast cancer, stroke, and pulmonary embolism risks for healthy postmenopausal women assigned to estrogen plus progestin (Writing Group for the Women's Health Initiative Investigators 2002). Commenting on an earlier

RCT, Petitti (1998, 650) had described similar findings as "a sobering reminder of the limitations of observational research, the incompleteness of current understanding of the mechanisms of vascular disease, and the dangers of extrapolation."

In 2008, Hernán and colleagues reported a reanalysis of the data from one of the large observational trials (the Nurses' Health Study) that had shown benefits from HRT. Their reanalysis, which "emulated the design and intention-to-treat analysis of the randomized trial" (Hernán et al. 2008, 766), concluded that differences between WHI and the Nurses' Health Study "could be largely explained by differences in the distribution of time since menopause and length of follow-up." The article by Hernán and colleagues (2008) was published with commentaries and an editorial, in which Wilcox and Wacholder (2008, 765) wrote, "[T]hese papers lay open a controversy that will not quickly be resolved. . . . The relation of randomized trials and observational studies has long been fertile territory for methodological discussion. Is this new reanalysis a distracting detour, or does it augur a change in the way we do our work?" Clearly, more empirical work is needed to clarify the nature and magnitude of differences associated with different analytic methods and how the use of various models can help reveal the underlying reality.

Notes

1. In "quasi-experiments," patients who receive an intervention are distinguished from those who do not by a weaker mechanism than randomization. For example, an intervention might be implemented earlier in one part of a large geographic area than in another as a way of learning about the intervention's effects. In contrast, in a purely nonexperimental observational study, the intervention is offered throughout the geographical area, and those who receive the intervention are compared to those who do not. For our purposes, we note that data from a quasi-experimental observational study resemble those from a nonexperimental observational study in that they both need to be adjusted for risk.
2. An "indicator" or "dummy" variable for a factor equals 1 when the factor is present and 0 otherwise. For example, in examining a population aged 65 or older, we could categorize people into seven age groups (65 to 69, 70 to 74, . . . , 90 to 95, and 95+) and enter six dummy variables that distinguish people in each of the other six groups from those aged 65 to 69.
3. If b were negative, women's expected costs would be lower. Notice also that in this model, the effect of sex is assumed to be the same

amount, b, for all categories of patients—for example, between male and female 30-year-olds and 60-year-olds. The way to avoid this restriction is to include interactions between gender and other risk factors—for example, by using age/sex indicator categories, such as 30- to 34-year-old females.

4. Weighted least squares is preferred if some kinds of people (e.g., members of a vulnerable minority) are more heavily sampled than others from a target population and the goal is to generalize findings to the whole population.
5. "Odds" are familiar from gambling, as in "the odds of winning are 5 to 2," or $5/2 = 2.5$. While probabilities must be between 0 and 1, odds can assume any positive value, and log odds can assume any value, negative, 0, or positive. For this reason, log odds are more attractive for modeling than raw probabilities are. The relationship between the probability of an event and the odds is straightforward: $p_i = O_i/(1 + O_i)$. For example, odds of $1/2$, 1, and 2 correspond respectively to log odds of -0.69, 0, and $+0.69$ and to probabilities of $1/3$, $1/2$, and $2/3$.
6. For people who do not receive the intervention, $I = 0$, making $e^{cI} = e^0 = 1$; in contrast, the multiplier for those who receive the intervention is e^c.
7. Given a set of model coefficients, the "likelihood function" for a set of observed dichotomous (0 or 1) outcomes is calculated by multiplying together the model-predicted probabilities for each of the observed cases. If there are n_0 cases for whom the outcome is 0 and n_1 cases for whom the outcome is 1, then $(n_0 + n_1)$ model-predicted probabilities are multiplied together. Roughly, the model fits best (and the likelihood function is largest) when the model predicts large probabilities (ps) for the cases for whom the event occurred and large values of $1 - p$ for the cases for whom it did not. A maximum likelihood method seeks the set of model coefficients that makes the product of these probabilities as large as possible.
8. To clarify the nature of the dependent variable in this model, first recognize that the probability that an event (e.g., death) occurs in an interval of time (t to $t + \Delta t$) is the product of two probabilities: first, the probability that death does not occur prior to time t, and second, the conditional probability that a person who has survived until time t dies prior to $t + \Delta t$. For example, consider 1,000 patients alive at time 0. In the first six months, 416 patients die (on average, 16 deaths per week over 26 weeks). Suppose that in the next week, 2 more patients die. Thus, in these data, the probability of surviving six months is 0.574 and, **among patients who were alive at the beginning of the week**, the probability of dying in the 27th week is 0.0035 (2/574).

Starting at time 0, the probability of dying in the 27th week is 0.002 (2/574 × 574/1,000). The hazard rate at six months, $h(6)$, is the limiting value of the probability of dying during the interval $6 + \Delta t$ as Δt is made smaller and smaller. In the same way that 0.002 equals the probability of dying in week 27, $h(6)$ × the probability of surviving six months equals the probability of dying "at" six months.

9. In a data set containing many more controls than cases, one may wish to retain two or even four times as many controls as cases in each propensity score quantile to improve the precision of the final estimate of treatment effect. Ratios larger than 4 to 1 do little to improve precision. In a data set in which few cases or controls fall into extreme quantiles, eliminating those quantiles from the analysis will prevent misleading extrapolations of estimates of treatment effectiveness to groups in which few people are observed.
10. The same goal is addressed by eliminating intervention cases in quantiles in which there are insufficient controls to match to all of the intervention cases.
11. If both the propensity score model and the multivariable regression are linear, and if the propensity score model includes the same risk factors the regression includes to predict outcomes, then including propensity score as an **additional covariate** has no effect on model predictions. If the model is nonlinear, predictions could be affected slightly.
12. The simplest form of density plot is a histogram rescaled to have area equal 1. More sophisticated density plots are smoothed versions of normalized histograms. See, for example, http://support.sas.com/documentation/cdl/en/grstatproc/62603/HTML/default/viewer.htm#a003155747.htm.
13. We used linear models to predict probabilities here for ease of explanation. Logistic models are generally viewed as more appropriate for predicting dichotomous outcomes. Often, however, simple linear models applied to 0/1 outcomes produce findings similar to those produced by logistic models. In the case of the simple models in this example, predictions from either type of model would be identical.
14. The conditions mean that U must be independent of the Xs separately among those who receive the intervention and among those who do not, although U is not required to be independent of the Xs in the combined (intervention + control) population.

CHAPTER

12

OMPARING OUTCOMES CROSS PROVIDERS

ene S. Ash, Michael Shwartz, Erol A. Peköz,
Amresh D. Hanchate

Risk adjustment facilitates meaningful comparisons of outcomes across groups of patients by accounting for differences in intrinsic patient characteristics that affect outcomes. Such comparisons, however, are ly a means to a larger end. Nightingale and Codman considered com- son of outcomes a powerful way to motivate hospitals to improve the ity of care (see Chapter 1). Nightingale (1863, 175–76) wrote:

> In attempting to arrive at the truth, I have applied everywhere for information, but in scarcely an instance have I been able to obtain hospital records fit for any purposes of comparison. . . . I am fain to sum up with an urgent appeal for adopting . . . some *uniform* system of publishing the statistical records of hospitals. There is a growing conviction that in all hospitals, even in those which are best conducted, there is a great and unnecessary waste of life . . .

htingale and Codman argued that simply comparing rates of events was fficient. One must discover why differences in patient outcomes occur correct identified problems.

Today, comparisons of outcomes are central to the scrutiny of the US lth care delivery system and cost containment strategies and are an impor- component of responses to competitive market forces. Patient outcomes compared across physicians; group practices; clinics, hospitals, and other itutional settings (e.g., nursing homes); and private and public health rers. These comparisons are variously called performance or practice files, report cards, scorecards, and outcomes reports. As noted in Chap- 1, the growing interest in pay-for-performance schemes will likely place n profiles center stage, heightening the financial stakes of risk-adjusted comes measures.

Several types of questions motivate report card and profiling initiatives— example:

Do any providers stand out as either much better or worse than average? How strong is the evidence that a certain provider's performance has been (or, perhaps more important, will be) substandard?

Numerous decisions must be made when designing a profiling approach and assembling data to compare patient outcomes across providers and answer such questions (Exhibit 12.1). In addition, interpretation of results depends on both good methodology and a thoughtful conceptual framework. This chapter discusses principles of good design and important practical considerations in performance profiling. We emphasize that no all-purpose best way exists to compare patient outcomes across providers. Especially when using profiles to support decisions with serious patient care or financial implications, analysts must be aware of how methodological choices may shape their findings.

Effects of Randomness on Comparing Patient Outcomes

Random fluctuations affect estimates of provider performance and thus limit the conclusions that can be drawn safely from performance profiles. To elucidate the role of randomness, we consider a contrived example using hospital costs as our outcome. We assume that the patients have identical clinical conditions and receive the same treatment across hospitals.

The simplest model views the costs of the n_A cases admitted to hospital A this year as a sample from a theoretically infinite population of cases that might be treated at hospital A. We will designate this year's observed average cost at hospital A as Y_A rather than as the more usual but cumbersome $\bar{Y}_A$.

EXHIBIT 12.1 Design Considerations for Provider Profiling

- What data will be used?
  - Can information be linked at the person level?
  - Can numerators and denominators be determined?
  - What are the accuracy and reliability of the data?
  - Which patient risk factors are captured in the data?
  - What is the time frame encompassed by the data?
- What outcomes can be measured from the data?
- Which providers will be included?
  - Are there reasons to exclude any providers?
    - Small sample sizes
    - Incomplete data
    - Known patient risks unable to measure with the data (e.g., public hospitals)
    - Policy considerations (e.g., small hospitals, rural hospitals)
- Which patients will be included?
  - What are the specific inclusion criteria (e.g., disease, surgery)?
  - Are there reasons to exclude any patients?

$\bar{Y}_A$ estimates the underlying (sometimes called "true") average cost, μ_A, for all cases that might be treated at hospital A. When referring to a generic average, we will continue to use $\bar{Y}$.

The distribution of observed costs for this year's patients provides information about the variability of costs among potential groups of patients at various hospitals. Examination of this distribution is always advisable. For example, analysts should look at standard summary statistics, including the mean, median, and standard deviation; minimum and maximum values; and values associated with different percentage points of the distribution (e.g., values demarcating the upper 1 percent and 5 percent of cases). Side-by-side box plots, sometimes called box-and-whiskers plots, can help us compare hospitals. Exhibit 12.2, which we discuss in detail later, shows box plots for hospital lengths of stay (LOS). Such plots help us identify likely errors (e.g., hospital stays with negative LOS or costs) or values that are correct but extreme (Tukey 1977). For instance, a hospital that has high average costs because all its cases were expensive differs from an institution in which one very expensive case raised its average costs by nearly $10,000 (e.g., one "million-dollar baby" among 100 average-cost newborns).

The *standard deviation* (*SD*) is the most common summary measure of variation for a variable *Y*. The *SD* for a population is designated by σ, and

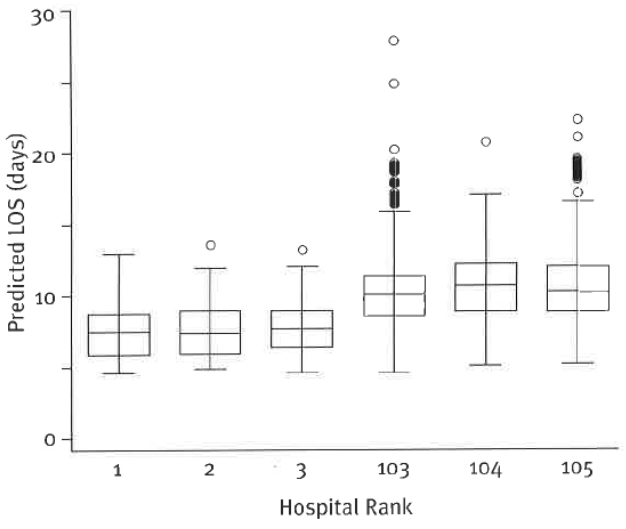


EXHIBIT 12.2 Box Plots of Expected Length of Stay (LOS) at Six Hospitals

Note: This exhibit includes box plots from three hospitals with the lowest and three with the highest expected LOS using Disease Staging's Relative Resource Scale to determine expected values.

its estimate, s, is computed for a sample $\{Y_1, Y_2, \ldots, Y_n\}$ from that population by taking the square root of:

$$s^2 = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{n-1}.$$

We use the sample mean, $\overline{Y}_A$, as our estimate of μ_A, the unobserved mean of Y in hospital A, and s_A as an estimate of σ_A, the *SD* of all outcomes of that theoretical population that could occur at hospital A. *SD*s are in the same units as Y (here, dollars), so Ys divided by *SD*s are unit-less. Thus, quantities of the form $(Y - \overline{Y})/s$ are said to be "standardized differences" and generally have values inside the range −2 to 2, or perhaps −3 to 3, because individual Ys usually lie within two or three *SD*s of their average.

The usual model hypothesized in comparisons of hospitals (or any other provider unit) assumes that

- each case at hospital A has an expected value μ_A, which may differ by facility; but
- the Y values at each facility are equally variable (i.e., there is a single, common value σ of the σ_As).

If the *SD*s in each hospital all have the same value, σ, each s_A (computed by applying the above formula to the cases at hospital A) is an estimate of σ. With K facilities $\{A_1, \ldots, A_K\}$, we combine the information about σ contained in the set of s_As to create a "pooled" estimate of σ, called s, by first estimating σ^2 (call it s^2) and then taking its square root. We compute s^2 as a weighted sum of the individual facility s_A^2s, using weights w_{A_k}s that sum to 1.0 and reflect differences in facility sample size:

$$s^2 = \sum_{k=1}^{K} w_{A_k} s_{A_k}^2,$$

where $w_{A_k} = (n_{A_k} - 1)/(N - K)$ and $N = \sum_{k=1}^{K} n_{A_k}$ is the total number of patients in all K facilities. If, for example, the n_{A_k}s were all equal, the w_{A_k}s would all equal $1/K$.

We often wish both to estimate the set of μ_As and to determine the accuracy of these estimates based on the Y_A, n_A, and s_A values. For most distributions of Y, and with fairly large sample sizes, Y_A is likely to fall within two standard errors (*SE*s) of μ_A—that is, within the interval $\mu_A \pm 2SE_A$, where SE_A, the (observed) *SE* of the mean at hospital A, is calculated as $SE_A = s/\sqrt{n_A}$. Thus, the intervals that go from $Y_A - 2SE_A$ to $Y_A + 2SE_A$ will contain most hospitals' mean value, μ_A.

Note that hospital A's central or mean value for Y (its μ_A) and how spread out its Y values are (its σ_A) are properties of the population that do

not depend on n_A. In contrast, Y_A is an estimate of μ_A based on n_A observations taken at hospital A. As sample size increases, SE_A decreases, reflecting the increased accuracy of Y_A as an estimator of μ_A. For example, suppose we observe mean costs of \$5,000 with an *SD* of \$5,000.[1] With 100 observations, we are reasonably sure that μ_A will be in the interval \$5,000 ± 2(500)—that is, between \$4,000 and \$6,000. With n = 400, μ_A is likely to be between \$4,500 and \$5,500.

Even with the highly skewed distributions typical of health care cost data (see Chapter 10), unless n_A is small, the observed mean tends to be approximately normally distributed around μ_A. This distribution implies that an interval centered at Y_A and extending two or three *SE*-sized units above and below will likely include μ_A. For many purposes, 30 cases is an adequate n_A. However, especially when the outcome variable is extremely skewed, hundreds of cases may be required to produce a nearly normal distribution of Y_A, as extreme values significantly affect the mean. When the underlying variable in the population from which the n_A cases were sampled is distributed normally, Y_A has about a 95 percent chance of falling in the "prediction interval," $\mu_A \pm 2SE_A$. In other words, the interval $Y_A \pm 2SE_A$ has a 95 percent chance of containing the true value μ_A. This latter interval is called the 95 percent confidence interval (CI). A CI specifies plausible boundaries for a parameter estimate (here, μ_A), whereas a *prediction interval* establishes boundaries within which an observed random variable (here, Y_A) should lie. Although cost distributions generally are not normal, cost averages, such as Y_A, are more nearly so.[2]

Profiling is not a theoretical exercise. Rather, these intervals help us identify real-world situations that may merit further examination. A two-*SE* cutoff casts a broad net to identify possible problems, many of which will be spurious. Three *SE*s might prove more appropriate for certain settings, such as public reporting, in which false flags have severe consequences.

For a positive random variable, it is useful to compute an observed *coefficient of variation* (CV), which expresses the variable's *SD* as a multiple of its mean:

$$CV = \frac{s}{\overline{Y}}.$$

The CV is a characteristic of the distribution of Y in the population; estimates of it do not get systematically smaller with larger samples. However, variables with large CVs require large samples to produce estimates of the mean that are usefully close to its true value. Exhibit 12.3 shows the half-widths of (approximate) 95 percent CIs:

$$2SE = 2Y_A \frac{CV}{\sqrt{n}},$$

for different values of CV and sample size. The figures in Exhibit 12.3 demonstrate the combined effect of sample size and data variability on how wide a CI must be to have a good chance of containing μ_A. For example, with CV = 1 and a sample size of 100, an approximate 95 percent CI for μ_A is $(0.8Y_A, 1.2Y_A)$—that is, we can estimate mean cost with about 20 percent error. Because accuracy is proportional to 1 divided by n, 400 observations are required to achieve estimates with 10 percent error. Depending on how similar the hospitalization admissions under consideration are, the CV for hospital cost, for example, can be much larger than 1.0 (Quinn 2008).

From a different perspective, assume that average costs are the same at each hospital. Then, because each μ_A equals μ, Y_A is likely to be within the interval $\mu \pm (2s/\sqrt{n_A})$.[3] If we assume a large number of patients across all hospitals, $\bar{Y}$ calculated from all patients is a good estimate of μ, the true mean cost for all patients. Hence, Y_A is likely to be within the interval $\bar{Y} \pm (2s/\sqrt{n_A})$. Suppose Y_A falls outside this interval. In the traditional hypothesis-testing framework, we conclude that hospital A's costs differ from the average. Furthermore, suppose that we are judging 100 facilities and flagging any hospital as an outlier when Y_A is outside the interval $\bar{Y} \pm (2s/\sqrt{n_A})$. In this situation, hospitals designated as outliers have costs that are statistically significantly different than the average at $p < 0.05$. Among 100 facilities performing identically (i.e., all μ_As = μ), this approach will, on average, incorrectly identify five as outliers. This example illustrates the "multiple

EXHIBIT 12.3 Effect of Sample Size and Coefficient of Variation (CV) on the Half-Width of Approximate 95 Percent Confidence Intervals*

	CV (σ/μ)					
Sample Size (n)	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**
10	0.32	0.63	0.95	1.26	1.58	1.90
25	0.20	0.40	0.60	0.80	1.00	1.20
50	0.14	0.28	0.42	0.57	0.71	0.85
100	0.10	0.20	0.30	0.40	0.50	0.60
150	0.08	0.16	0.24	0.33	0.41	0.49
200	0.07	0.14	0.21	0.28	0.35	0.42
300	0.06	0.12	0.17	0.23	0.29	0.35
400	0.05	0.10	0.15	0.20	0.25	0.30
500	0.04	0.09	0.13	0.18	0.22	0.27
1,000	0.03	0.06	0.09	0.13	0.16	0.19

*Cells of the table are $(2 \times CV)/\sqrt{n}$. Half-width = $Y_A \times$ table cell.

comparisons" problem (Snedecor and Cochran 1980). Incorrectly flagging a nonproblematic hospital is called a type I error.

On the other hand, trying to avoid type I errors by being conservative about flagging outliers increases the chance that true outlier hospitals are missed—a type II error. For an illustration of type II errors, assume that an inefficient hospital (hospital I) has costs 20 percent above an average of $1,000. This difference seems sufficiently large to be important. Suppose we flag a hospital as an outlier only when its observed mean differs from $1,000 by at least 40 percent (roughly the cutoff for identifying a statistically significant difference at $p < 0.05$ when $n = 100$ and the CV = 2). Under this rule, hospital I will be flagged if its average cost exceeds $1,400 (i.e., the 95 percent CI for μ_I lies entirely above $1,000). However, with only 100 observations and a CV of 2, the chance that hospital I's observed mean will exceed this threshold is only 20 percent. Thus, hospital I has an 80 percent chance of avoiding an outlier flag.

The same considerations apply when we examine a dichotomous outcome, such as death. Assume that D is the death rate in a large population of patients and d is the observed rate (the proportion who died) in a smaller population of size n (e.g., patients at hospital A). The estimated SE is then $\sqrt{d(1-d)/n}$. For different values of n and death rate D, Exhibit 12.4 shows the half-width of an interval that is about 95 percent likely to contain D.[4] For

Sample Size (n)	Probability of Death (D)**							
	0.01	0.02	0.05	0.10	0.15	0.20	0.25	0.50
25	—	—	—	0.12	0.14	0.16	0.17	0.20
50	—	—	0.06	0.08	0.10	0.11	0.12	0.14
100	—	—	0.04	0.06	0.07	0.08	0.09	0.10
150	—	—	0.04	0.05	0.06	0.07	0.07	0.08
200	—	—	0.03	0.04	0.05	0.06	0.06	0.07
300	—	0.02	0.03	0.03	0.04	0.05	0.05	0.06
400	—	0.01	0.02	0.03	0.04	0.04	0.04	0.05
500	0.01	0.01	0.02	0.03	0.03	0.04	0.04	0.04
1,000	0.01	0.01	0.01	0.02	0.02	0.03	0.03	0.03

EXHIBIT 12.4 Effect of Sample Size and Probability of Death on the Half-Width of Approximate 95 Percent Confidence Intervals*

*Cells of the table (= half-width) are $2\sqrt{D(1-D)/n}$.

**When $n * D$ (the expected number of deaths) < 5, the normal approximation (the basis for calculations in this table) is unreliable.

Note: More precise CIs for proportions are described in Agresti and Coull (1998) and implemented in www.graphpad.com/quickcalcs/ConfInterval2.cfm.

example, if, in a sample of 100 patients, d were 10 percent, the interval from 4 percent to 16 percent would likely contain D.[5] In many settings, this interval is wide compared to reasonable differences between poor- and high-quality providers (Hofer and Hayward 1996; Ash 1996).

With any rule for flagging outliers, as the difference increases between a given hospital's underlying performance and typical performance, the likelihood of being flagged rises. Thus, depending on the (unknown) mix of normal and variously aberrant providers in a study population, roughly 5 percent of nonproblematic providers will erroneously receive outlier flags, whereas some (unknown fraction of) problematic providers will escape flags. Which flags are incorrect is generally not obvious. Using data on cardiac catheterization, Luft and Hunt (1986, 2780) illustrated that small numbers of patients and relatively low rates of poor outcomes make it difficult to "be confident in the identification of individual performers." For example, suppose the death rate is 1 percent, but a hospital treating 200 patients experiences no deaths. Even when we use a lenient 0.10 significance level, determining whether that hospital had statistically significantly better outcomes is impossible. If the expected death rate is 15 percent and five deaths occurred out of 20 patients (an observed rate of 25 percent), the difference is insufficient to label the hospital a poor performer. Thus, random chance plays a prominent role in determining outlier status when sample sizes are relatively small. In this situation, comparisons across providers must be interpreted cautiously.

Comparing Observed and Expected Outcomes

Calculation of expected rates of outcomes is usually the first step in producing risk-adjusted performance profiles. The simple example just presented ignores the need for risk adjustment by considering patients with identical clinical conditions. In most situations, different providers see different mixes of patients, so risk adjustment is essential.

Linear regression modeling is the most commonly used method for risk-adjusting continuous outcomes (see Chapter 10). Thus, we might build a model as follows:

$$PRED_i = a + \sum_{j=1}^{J} b_j X_{ij},$$

where $PRED_i$ is the expected outcome for patient i, who has characteristics X_{ij}, for the J predictors in the model. The expected mean outcome for patients treated by a specific provider (E) equals the average of their $PRED_i$s.

In contrast, logistic regression, in which the log of the odds of the event is modeled as a linear function of the predictor variables, is generally

used to predict dichotomous (yes/no) outcomes.[6] After fitting a logistic regression model, we calculate the predicted probability of death for the i^{th} case ($PRED_i$) from the relationship

$$ln(odds)_i = ln\left(\frac{PRED_i}{1 - PRED_i}\right) = a + \sum_j b_j X_{ij}$$

by solving for $PRED_i$:

$$PRED_i = \frac{e^{ln(odds)_i}}{1 + e^{ln(odds)_i}}.$$

To determine the expected number of deaths in a group of n cases, we sum the $PRED_i$ terms; to determine the expected death rate, we divide this sum by n.

Comparison of observed to expected outcomes is central to performance profiling (e.g., drawing inferences about the quality or efficiency of care). Various approaches have been used to compare O and E. Neither $(O - E)$ nor O/E is clearly a superior way of characterizing the difference. For example, suppose that hospital A treats cases expected to average \$5,000 (i.e., average $PRED_i$), but the actual cost is \$6,000. In contrast, hospital B treats cases that should cost \$10,000 but actually average \$11,500. Thus, both hospitals' costs are greater than expected, but how do they compare with each other? On an additive or difference scale $(O - E)$, hospital B performs worse, as hospital A's excess is only \$1,000 per case, compared to hospital B's excess of \$1,500. However, on a multiplicative or ratio scale (O/E), hospital A does worse, as its cases cost 20 percent more than expected whereas hospital B's cost 15 percent more. Theory offers no insight into which hospital performs better.

Consider another example. Which is worse: 2 percent complications when only 1 percent was expected (double the rate), or 50 percent complications when only 40 percent was expected (10 excess problems per 100 but only a 25 percent higher complication rate)? Mathematics cannot answer this question. Analysts can use their data to explore which is more realistic: an additive model (in which adding the same amount to each case represents the provider effect) or a multiplicative model (in which provider-associated increases are proportionate to the expected outcome). Even when multiplicative models are chosen, observers typically still want to know how observed results compare additively to expected results—for example, how many extra dollars a provider costs or how many extra complications have occurred.

Ratios, such as O/E, are centered at 1 but range from 0 to infinity. To put comparable distances between ratios below 1 and those above 1, analysts sometimes display $\ln(O/E)$ values rather than O/E values (Roos, Wennberg, and McPherson 1988). On a graph displaying O values on the x-axis and $\ln(O/E)$ values on the y-axis, a "broken" y-axis can be used to indicate the gap between the smallest $\ln(O/E)$ associated with a positive observed (O) and

negative infinity (which is the value of the logarithm function at zero). On an untransformed scale, substantial differences among *O/E* values less than 1 are difficult to see and thus may appear unimportant. In contrast, on a log scale, the distances between points representing *O/E* values of 0.25, 0.50, 1.00, 2.00, and 4.00 are proportionately spaced because each value doubles the one below it.

A drawback of the ratio *O/E* is that when *E* is small, its value changes dramatically with small changes in *O*. For example, if we observe 30 cases, each with a 1 percent risk of complications, the expected number of complications is 0.3. If 0, 1, or 2 complications are observed, *O/E* is 0, 3.3, or 6.7, respectively. A good guideline is to avoid examining such ratios when the expected number of events is less than 1.0. Some researchers advise against *O*-to-*E* comparisons unless *E* is at least 5.

Fortunately, when comparing *O* to *E*, findings as extreme as our examples are unlikely. If expected costs at two hospitals are $5,000 and $10,000, or if expected complications rates are 1 percent and 40 percent, their patient populations or other characteristics probably differ too much for comparison to be useful; when distributions of expected outcomes are roughly similar across hospitals, difference and ratio measures of performance produce reasonably similar results. Examining *expected outcomes* across providers is therefore important to ensure that patients' risks do not differ radically across providers (see the following paragraph). A useful first cut at comparing expected outcomes across providers is to review common descriptive statistics (e.g., mean and median, *SD*, percentage cutoffs of the distribution, box plots).

In our past work, we examined the extent to which severity explained differences in hospital LOS for pneumonia patients (Iezzoni et al. 1996c). To help us understand which hospitals could reasonably be compared, we examined the distribution of expected LOS at each hospital. (We determined the expected values by using the Disease Staging Relative Resource Scale. We removed outliers using Medicare's definition at the time: cases with costs more than three *SD*s above the mean on a log scale; see Chapter 10.) To illustrate differences in these distributions, consider six hospitals (the number of cases per hospital ranged from 73 to 316): three hospitals with the highest and three with the lowest expected average LOS. Exhibit 12.2 shows side-by-side box plots of expected LOS values at these six hospitals.

In Exhibit 12.2, the box shows the range encompassing the middle 50 percent of cases. Thus, 25 percent of cases have values below the bottom edge of the box, and 25 percent have values above the upper edge. The horizontal line within the box is the median. The length of the box is the *interquartile range* (IQR), sometimes called the *H-spread*. The top of the box plus 1.5 IQR and the bottom of the box minus 1.5 IQR define the inner fences; the top of the box plus 3 IQR and the bottom of the box minus 3 IQR define

the outer fences. The ends of the lines extending above and below the box indicate the highest and lowest values within the inner fences; circles indicate individual values between the inner and outer fences. Different computerized statistical packages use different symbols, but the box plot concepts are similar. The box plots show that 75 percent of patients in the three hospitals with the lowest expected LOS were expected to have an LOS of less than eight days, whereas 75 percent of cases at the hospitals with the highest expected LOS were expected to have an LOS of greater than eight days.

Failure of *O*-to-*E* Comparisons to Adjust Fully for Risks

When examining death rates, epidemiologists often use standardized mortality ratios (SMRs). SMRs are *O/E* ratios whose *E* values are calculated using indirect standardization. To better understand the need for standardization (epidemiologists' term for risk adjustment), consider a hypothetical situation involving two types of patients: low risk, with a 1 percent mortality rate, and high risk, with a 5 percent mortality rate (Exhibit 12.5). Suppose that half of all patients in a large population are low risk and half are high risk, yielding an overall mortality rate of 3 percent. Now consider hospital A, which treats 1,000 patients, 800 of which are low risk and 200 high risk. Hospital A's experience with its low-risk patients is similar to the overall experience: a 1 percent mortality rate (8 deaths among the 800 patients). However, hospital A does poorly with high-risk patients; it has a 10 percent mortality rate, double the population average (20 deaths among the 200 high-risk patients). Despite this poor performance, because of its favorable case mix hospital A's mortality rate is 2.8 percent (28/1,000), somewhat better than the 3 percent population average.

	All Patients in Population		Hospital A		Hospital B		Hospital C	
Risk Category	Patient Mix (%)	Death (%)	*n*	Death (%)	*n*	Death (%)	*n*	Death (%)
Low	50	1	800	1	200	1	800	1.25
High	50	5	200	10	800	10	200	12.50
Performance								
Observed death rate (*O*)		3	28/1,000 = 2.8%		82/1,000 = 8.2%		35/1,000 = 3.5%	
Standard mortality ratio (SMR) = *O/E*			28/18 = 1.56%		82/42 = 1.95%		35/18 = 1.94%	
Risk-adjusted mortality			3 * 1.56 = 4.68%		3 * 1.95 = 5.85%		3 * 1.94 = 5.82%	
Difference (*O* – *E*)			2.8 – 1.8 = 1%		8.2 – 4.2 = 4%		3.5 – 1.8 = 1.7%	

EXHIBIT 12.5 Hypothetical Hospitals with Different Patient Mixes and Death Rates

Indirect standardization determines a hospital's expected number of deaths by applying stratum-specific rates determined from all patients to the number of cases in each stratum in the hospital. In this case, a stratum is a risk category. On the basis of the overall data, we expect 8 deaths among the 800 low-risk patients (with a 1 percent mortality rate) and 10 deaths among the 200 high-risk patients (with a 5 percent mortality rate), for an expected rate of 1.8 percent.[7] The *O/E* ratio for hospital A is 1.56 (28/18) because it has 56 percent more deaths than expected for its patient mix.

One can report this discrepancy in other ways. Some prefer to express the hospital's performance on the same scale as the population average, producing a standardized mortality ratio (SMR) by multiplying the *O/E* by the population average rate (e.g., 1.56 × 3 = 4.68 percent). Another choice is to report the difference between the observed rate (2.8 percent) and the expected rate (1.8 percent). Thus, hospital A has 1 percent more deaths than expected. These summary measures all agree on the main point: After adjusting for its patient mix, hospital A has more deaths than expected.

Indirect standardization and its generalization via multivariable risk adjustment modeling are powerful tools for making fairer comparisons among providers with different types of patients. Nevertheless, comparing outcomes across providers is complicated when patient mix both strongly affects the outcome and differs widely across providers. In the terminology of epidemiology, patient mix is a confounding factor in examination of patient outcomes.

For an illustration, consider hospital B, which has the same mortality experience within each stratum as hospital A, but an unfavorable case mix. Hospital B treats 200 low-risk patients, resulting in 2 deaths, and 800 high-risk patients, resulting in 80 deaths (see Exhibit 12.5). Solely because of differences in patient mix, hospital B's unadjusted death rate is 8.2 percent, much higher than hospital A's 2.8 percent rate. When facilities' patient mixes differ widely, "raw" comparisons can mislead.

However, risk adjustment does not always do what we anticipate or hope. For example, an indirect adjustment approach fails to make hospitals A and B look equally good. To perform indirect adjustment for hospital B, we first compute its expected number of deaths as 42 ([0.01 × 200] + [0.05 × 800]). Hospital B's SMR is thus 1.95, its risk-adjusted death rate is 5.85 percent, and its excess mortality rate is 4 percent (as opposed to 1.56 percent, 4.68 percent, and 1 percent, respectively, at hospital A). However reported, hospital B looks worse than A, although the same type of patient had the same outcome at both hospitals. Results could be even more misleading. Imagine that hospital C is seriously deficient: It has the same favorable patient mix as hospital A but 25 percent higher death rates for both patient types (1.25 percent and 12.5 percent mortality, respectively, among low- and high-risk patients). Hospital C's SMR, risk-adjusted death rate, and excess

mortality rate (1.94 percent, 5.82 percent, and 1.7 percent, respectively) look marginally better than hospital B's, although its performance is clearly worse.

Direct standardization, an alternative adjustment approach, produces results that feel more correct, but the method presents conceptual and practical problems. In direct standardization, provider-specific rates are computed in each risk stratum and applied to a "standard" population case mix, producing an estimate of what might be expected if the provider were to treat this standard patient mix. For example, suppose that the standard population has 50 percent low-risk and 50 percent high-risk patients. Under this assumption, hospitals A and B have stratum-specific death rates estimated to yield 5.5 percent mortality in the standard population ($[0.5 \times 0.01] + [0.5 \times 0.10]$), compared to hospital C's estimated 6.9 percent rate ($[0.5 \times 0.0125] + [0.5 \times 0.125]$).

In epidemiological studies, the strata are generally large (e.g., populations in different states are broken down into five-year age categories). Relatively reliable estimates of stratum-specific rates are possible using such large populations. However, when one profiles individual providers for patients stratified by disease or other risk factors, stratum-specific rates are generally based on too few cases to provide reliable estimates. Furthermore, questions arise about whether a provider should be judged harshly for doing poorly with types of patients it rarely sees. For example, suppose hospital D treats 1 high-risk patient who dies and 999 low-risk patients, of whom only 5 die. Although its death rate is only half as large as the 1 percent rate expected for nearly all of its 1,000 patients, its projected death rate for the standard population is more than 50 percent ($[0.5 \times 0.005] + [0.5 \times 1.00]$). Thus, as this example demonstrates, which of several providers looks best can change depending on the patient mix of the standard population. Direct standardization is rarely used to profile physicians or hospitals.

Most performance profiles use more complex multivariable models to determine expected values. However, the fundamental approach is identical: Each provider's observed outcome is compared to expected outcomes based on the risk characteristics of its patients and the model-specified relationship between these characteristics and the outcome of interest. When providers treat very different populations, therefore, risk adjustment cannot definitively identify which performs better. In reality, particular providers may do better with certain types of patients and worse with others. Thus, examining the data in multiple ways (e.g., examining providers within high- and low-risk strata of patients) is important. In a rational world, providers would concentrate on their most successful types of cases, and performance profiles would steer patients to providers who do well with these kinds of patients. Shahian and Normand (2008) caution against direct comparison of two institutions unless the types of patients they treat substantially overlap.

Random Variation in Comparing *O*-to-*E* Outcomes

As discussed, standard errors capture the effect of random variation on the reliability of estimates from data. When each of n observations is an independent observation from a common distribution with mean μ and σ, we estimate μ by $\overline{Y}$ and σ by:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{n-1}}.$$

The values of $\overline{Y}$ and s remain relatively constant as n increases, becoming increasingly accurate estimators of μ and σ, respectively. In contrast, the statistic that measures the error associated with using $\overline{Y}$ as an estimator of μ decreases as n increases: $SE(\overline{Y}) = s/\sqrt{n}$.

Analysts must be careful when estimating a denominator to help them interpret the size of the difference between an O and an E in a single facility (i.e., creating a standardized difference for the facility). Consider predictions of a continuous outcome from a multivariable linear regression model. Most computerized statistical regression packages provide two distinct *SE* values associated with an observed Y. One relates to the size of the error associated with calculating $\hat{Y}$ or *PRED* as an estimate of its true expected value, $\mu_{\hat{Y}}$, which is close to zero when the model is assumed to be correct and has been fit to a very large data set; the other relates to the (much larger) variability of a single observation Y around *PRED*. For provider profiling, the latter *SE* is relevant because variability in the observed outcome from a modest-sized sample of Ys contributes almost all the "noise" in $(O - E)$.

Let s_i be the larger standard error for the i^{th} observation output by the regression software. An approximate *SE* and 95 percent CI for the true average μ_A based on the n observations at hospital A can be calculated as:

$$SE = \frac{\sqrt{\sum_{i=1}^{n} s_i^2}}{n} \text{ and } Y_A \pm 2SE.$$

The distributions of continuous outcomes, such as costs and LOS, usually have "long right tails," including some cases with extremely high values (see Chapter 10). Therefore, the logarithm (usually the natural logarithm) of these continuous values is often used in modeling because its distribution is more symmetrical than that of the untransformed data. In this situation, CIs can be computed on the log scale. However, achieving estimates on the original scale requires that the point estimate of the mean and the endpoints of the CI be retransformed by exponentiation (and adjusted for bias; see Chapter 10). Resulting CIs will not be centered at the estimated mean.

Consider a dichotomous outcome, such as death. If person i's probability of dying is D_i, the *SD*s and *SE*s associated with (1) observing whether

person i dies, (2) the number of people dying among n persons, and (3) their mean death rate are, respectively:

$$s_i = \sqrt{D_i(1 - D_i)},\ s = \sqrt{\sum_i s_i^2},\ \text{and}\ \frac{s}{\sqrt{n}}.$$

We do not know the D_is, so we use each person's estimated probability of dying (from our model) in their place.

When each provider's expected outcome (E) is determined from a model fit to many cases, calculations of 95 percent CIs for O/E ratios generally treat E as a constant, resulting in an (approximate) 95 percent CI of $(O \pm 2SE)/E$.

Multiplying each end of the CI by the area-wide rate yields a CI for risk-adjusted outcomes. Hosmer and Lemeshow (1995) found that this approach worked fairly well in simulation studies.

The above CIs are correct (i.e., have approximately 95 percent coverage properties) when the observed number of cases follows a binomial distribution. When calculating CIs, analysts sometimes assume that the observed number of cases follows a Poisson distribution. New York State, for example, assumes a Poisson process in its published profiles of facilities performing coronary artery bypass graft (CABG) (Racz and Sedransk 2010). Their CIs are based on Byar's approximation of the exact Poisson distribution, which is accurate even with small numbers (see Breslow and Day 1987, 69). For an O/E ratio, this approximation leads to lower and upper 95 percent CIs calculated as follows:

$$Lower\ bound = \frac{O}{E}\left[1 - \frac{1}{9O} - \frac{1}{\sqrt{3O}}\right]^3$$

and

$$Upper\ bound = \frac{O}{E}\left(\frac{O+1}{O}\right)\left[1 - \frac{1}{9(O+1)} - \frac{2}{\sqrt{3(O+1)}}\right]^3$$

Bayesian Models

Standard approaches to provider profiling present several problems. The first is how the "true" mean value of an outcome is estimated for each provider (e.g., μ_A for provider A). Traditionally, it is calculated separately for each provider as the average outcome of patients treated by that provider (Y_A). However, the resulting set of provider estimates is often not as close as it could be to the true means and not the best predictor of what will happen in

the future. Typically, Y_As are too extreme, the highest ones higher than the associated true μ_As and the lowest Y_As too low. When provider-specific averages are based on small numbers of patients, they may be greatly over- or underestimated.

Traditional estimates of *SE* (described earlier) may understate the total variability present in the data, resulting in overly narrow CIs and too many average providers flagged as outliers. One reason for this understatement is that these methods recognize only the variation of individual patient outcomes around their *PRED*s. However, variation also occurs among providers. In addition, *SE*s are often underestimated because patients treated by a particular provider are likely to be more similar to each other (for whatever reason) than to patients treated by another provider. In other words, patients may not be independent observations but rather "clustered," and physicians may be "nested," often by some organizational hierarchy. For example, patients are clustered by their treating physicians, and physicians are often nested within hospitals or group practices. When one analyzes units within which clustering occurs, effective sample sizes (in terms of the amount of information provided) are smaller than actual sample sizes. Approaches that do not adjust for clustering underestimate *SE*s (Greenfield et al. 2002). Hierarchical models, also called *multilevel* or *random coefficient* models, take a comprehensive approach to dealing with such problems.[8]

The Bayesian Approach

Approaching hierarchical models from a Bayesian perspective, which views newly observed data within the context of prior knowledge, has conceptual advantages. For example, suppose in ten coin tosses we observe one head. We know from our understanding of coin tosses that *PROB* (the true probability that a coin comes up heads) is close to 0.5. Therefore, our observed rate of 0.1 heads makes us question whether the coin is truly fair.[9] However, if the coin looks normal, it still seems more reasonable to suppose that the *PROB* is closer to 0.5 than 0.1. If we toss the coin again and continue to generate fewer than 50 percent heads, we grow more suspicious that *PROB* is less than 0.5. If the coin is biased, what is its true *PROB*? As the number of coin tosses increases, the observed proportion of heads increasingly becomes a more credible estimate of *PROB*.

Similar reasoning is useful for evaluating providers such as hospitals, in which case, without data, we assume that hospital A has outcomes like those at other hospitals. (This premise is called the *exchangeability assumption*.) Upon seeing a few data about hospital A's performance, we adjust this initial estimate slightly. As evidence accumulates, we place increasing weight on the new data and less on our initial beliefs. At some point, enough data might exist to convince us that hospital A's quality or efficiency is truly lower (or higher) than that of other hospitals.

Bayesian analyses interpret new data within the context of prior (initial) beliefs. For another example, if we are uncertain about the safety of an operation but observe one complication in ten operations, 0.1 might be a reasonable guess for the true complication rate, P_{COMP}, especially if similar surgeries have had 10 percent complication rates. However, if we observed no complications in ten operations, we may feel uncomfortable with $P_{COMP} = 0$ because we know that all surgery presents risks. In either case, we should recognize that ten are too few observations from which to be able to say much about the true rate.

Classical statistical methods capture the level of uncertainty in estimating P_{COMP} at a particular facility by putting CIs around its estimated value, as described earlier. These computations rely only on observations from that facility (usually within a single year) and may lead to "all-or-nothing" judgments—that is, if a 95 percent CI does not include the expected value, we accept the facility's observed mean as our best estimate of its true P_{COMP}. Otherwise, we are supposed to accept, at least implicitly, that the facility's P_{COMP} is best estimated from general knowledge, such as the average rate across all facilities or some value previously viewed as reasonable. For example, observing only one head in ten coin tosses of an apparently fair coin (i.e., coin with *PROB* = 0.50) leads us to estimate *PROB* as 0.10—after rejecting, at the $\alpha = 0.05$ level, our prior belief that the coin is likely to behave as other fair coins do. However, had our observed rate of heads been between 20 percent and 80 percent, we would have accepted that the coin is fair—that is, our best guess for *PROB* would remain 0.50.

A Bayesian framework uses prior knowledge about a situation to produce estimates for the true mean that lie somewhere between the observed average and the expected mean based on prior knowledge. The resulting estimate is closer to the expected mean when the observed mean is based on few data and prior knowledge (e.g., about how apparently fair coins behave) is strong. The estimate is closer to the observed mean when either outside knowledge is less certain (e.g., surgery with an unknown complication rate) or more data are available (e.g., when the observed mean derives from 1,000 operations rather than 10).

Bayesian thinking—starting with an expectation based on accumulated wisdom from observations of similar phenomena and gradually updating one's thoughts about a situation based on situation-specific data—is natural. However, Bayesian dependence on prior knowledge is controversial: Two people examining identical data could reach different conclusions as a result of having different assumptions about prior knowledge (i.e., "priors"). Fortunately, analysts can use a Bayesian framework without depending on external priors. One such approach is called *empirical Bayes analysis*, in which the whole data set is used to establish general knowledge about how an individual facility is likely to perform, while facility-specific data are used

to particularize this knowledge. Another strategy is to assume vague prior knowledge, which is captured by placing noninformative prior probability distributions on unknown parameters (e.g., by conjecturing that complication rates associated with a new surgery are uniformly distributed between 0.1 percent and 80.0 percent). When prior knowledge is vague, data primarily drive Bayesian estimates. Today, Bayesian analysis is considered mainstream statistics, not only because of its intuitive appeal and demonstrated value but also because today's powerful computers make the required computations feasible.

Empirical Bayes Analysis

Casella (1985, 83) attributes the basis of "modern" empirical Bayes analysis to work by Efron and Morris (1972, 1973, 1975). In a nontechnical paper, Efron and Morris (1977) show how parametric empirical Bayes analysis derives from a theorem of Stein (1956), who proved that one can better estimate the means of three or more normal populations jointly than by using the three averages computed separately from samples in each population. Exploration of the implications of Stein's theorem has led to even better ways to estimate means for several populations simultaneously.[10]

Empirical Bayes estimation uses information from the current data set as priors. For instance, in a database of 10,000 acute myocardial infarction (AMI) patients treated at 100 hospitals, 13 percent die in hospital. This 13 percent is used as prior knowledge; it provides a context for interpreting hospital A's experience (10 deaths out of 100 patients). The empirical Bayes estimate for the true death rate at hospital A will lie between 10 percent (situation-specific observed estimate) and 13 percent (prior-knowledge surrogate). Exactly where the empirical Bayes estimate falls depends on the relative size of estimates of random variation in these two numbers.

For an illustration of the empirical Bayes approach, consider comparing costs at four hospitals. A typical classical analysis considers two alternatives: (1) accept the null hypothesis, in which case the true mean cost at each hospital is estimated as the common mean from the pooled sample of all patients, or (2) reject the null hypothesis, in which case the mean of each hospital is estimated as the average of patients in that hospital. The empirical Bayes estimator is a compromise, estimating each hospital's mean by giving weight to both the common mean and the mean at each hospital. Thus, the empirical Bayes estimate of the average cost in each hospital "shrinks" the hospital-specific cost toward the overall average.

Empirical Bayes estimates explicitly recognize two sources of variation in the data: (1) random variation within each unit examined (e.g., within each hospital, variation of individual patients' observed costs from the hospital's true average cost, measured by σ_A for hospital A) and (2) variation across hospitals' true average costs (i.e., variation in the μ_i values for $i = 1$ to N, the

number of hospitals). In empirical Bayes analysis, the weight given to the observed mean in each unit is a function of these two sources of variation, measured by the variance (which equals the standard deviation squared):

$$weight = \frac{variance\ across\ units}{variance\ across\ units + variance\ within\ units}.$$

As variation within units (e.g., hospitals) increases, unit-specific averages receive less weight (i.e., estimates shrink closer to the overall average). As noted earlier, variation in the average is estimated to be $s/\sqrt{n}$, making within-unit variation larger in smaller samples. Usually, the most extreme raw averages come from units with small sample sizes. Thus, their Bayes estimates are much closer to the overall mean, leaving the most extreme Bayes estimates for units with less extreme raw averages based on larger samples.

We used empirical Bayes techniques to profile small geographical areas on the basis of hospitalization rates among persons aged 65 or over in Massachusetts (Shwartz et al. 1994). Specifically, we examined "relative hospitalization rates" (*RHR*) in each geographical area, defined as the observed number of hospitalizations minus the expected number, expressed as a multiple of expected: $RHR = (O - E)/E$. Thus, for example, *RHR*s of -0.5, 0.0, and +0.5 represent areas with 50 percent fewer hospitalizations than expected, as many as expected, and 50 percent more than expected, respectively. We determined expected numbers of hospitalizations using indirect standardization to adjust for differences in age and sex distributions of the population in each area. Empirical Bayes shrinkage, for example, affected perceptions of hospitalizations for cardiac catheterization. The highest *RHR* for cardiac catheterization, 0.90, occurred in a very small area with only 4,955 residents over age 64. The second-highest cardiac catheterization *RHR*, 0.84, came from a much larger area, with 40,390 residents over age 64. The empirical Bayes estimates for the two areas were 0.65 and 0.80, respectively. Because the first area had a small population, the empirical Bayes estimate gave less weight to its observed rate and more to the overall mean of 0. In other words, the rate estimated by empirical Bayes shrank much closer to the overall mean, from 0.90 to 0.65. Because the second area had a much larger population, far less shrinkage occurred. This example illustrates how empirical Bayes techniques adjust point estimates to reflect the uncertainty associated with raw averages, helping analysts guard against drawing conclusions from extreme estimates based on a few cases.

Our study also found that the set of empirical Bayes estimates of hospitalization rates in small geographical areas generally matched the set of area-specific rates for the following year better than did the raw averages (Shwartz et al. 1994). For 62 of the 68 conditions studied, empirical Bayes estimates yielded smaller weighted average errors (weighting by the size of the areas) when used to predict next year's hospitalization rates.

Note that a shrunken (Bayes) estimate may be worse than the average of the data for some units. Across all units, however, shrunken estimates produce lower overall errors.

Hierarchical Models

Hierarchical models generalize the idea of shrinkage and provide a comprehensive framework for explicitly incorporating variation at different levels of analysis (Bryk and Raudenbush 2002; Snijders and Bosker 1999). The "hierarchy" derives from nesting, which occurs when the data are not generated independently but in groups. For example, patients are nested within providers (e.g., groups of patients treated by the same physician), providers may be nested within practice groups (e.g., physicians who work at the same hospital), and hospitals may be nested within types (e.g., teaching versus nonteaching). At each level of the hierarchy, the relevant independent variables and their influence may differ. Explicit modeling of the hierarchical structure recognizes that nested observations may be correlated and that different sources of variation can occur at each level.

Consider the cost of treating patients with a particular disease; to simplify, assume patients are clinically similar across providers. To profile provider costs, we could use the following hierarchical model:

- Y_{ij} has some distribution (e.g., normal distribution truncated to be positive) with mean μ_j and variance σ^2 (stage I model).
- μ_j has some distribution (e.g., normal distribution truncated to be positive) with mean λ and variance v^2 (stage II model).

Y_{ij} is the observed outcome for person i treated by provider j. Provider j's true expected outcome is μ_j. We assume that random variation of outcomes is identical for each provider; variation is measured by σ^2. However, in reality, providers' true expected outcomes, the μ_js, are likely to differ. We assume that they follow a normal distribution with mean λ and variance v^2. Hence, we can generate a data point by (1) randomly selecting a μ_j from a positively truncated normal distribution with mean λ and some variance v^2 and (2) randomly selecting a Y_{ij} from a positively truncated normal distribution with mean μ_j and variance σ^2. Thus, the Y_{ij} values have two sources of variation, one due to variation within provider (σ^2) and another due to variation across providers (v^2). In a Bayesian hierarchical framework, the stage II parameters λ and v^2 are called *hyperparameters*. They can be given noninformative prior distributions (which implies little prior knowledge). For example, distributions can be specified as uniformly distributed over a wide range that incorporates all feasible values. The use of noninformative priors

allows the data (rather than information external to the data set) to primarily determine estimates. People who are uncomfortable with Bayesian models can model hierarchical structures within a strictly non-Bayesian framework, using maximum likelihood (ML) or residual (or restricted) maximum likelihood (REML) techniques to estimate parameters (Snijders and Bosker 1999; Raudenbush and Bryk 2002). Analysts can easily extend the simple model we just provided by incorporating individual-level risk factors or a risk score in the stage I model and provider-level covariates (e.g., physician specialty, practice site) in the stage II model.

Hierarchical models have several key features (Thomas, Longford, and Rolph 1994):

- They explicitly model differences among providers (beyond variations explained by differences in patient mix).
- They consider provider effects to be "random variation"; the measure of spread, v^2, is estimated during model fitting.
- They shrink the point estimate of a provider's outcome from the observed provider average toward a risk-adjusted expected value for the provider by an amount that depends on v^2, σ^2, and the provider's sample size.
- They produce wider intervals around point estimates that appropriately reflect the uncertainty arising from both individual variation of patients within providers and variation across providers.
- They provide a framework for comprehensively addressing the problem of multiple comparisons.

Gatsonis and colleagues (1995) offer a good nontechnical illustration of hierarchical modeling in examining variations across states in the use of coronary angiography for more than 218,000 elderly AMI patients. Patients were nested within states; states were nested within regions. In stage I (level I), Gatsonis and collaborators modeled for each state the probability that a patient received angiography, using logistic regression as a function of patient age, sex, race, and comorbidities. The researchers coded independent variables so that the intercept in the state was the log odds that a baseline case (65-year-old nonblack man with no comorbidities) received angiography. In stage II, they modeled intercepts from the stage I models as a function of region and a measure of the availability of angiography in the state. They developed stage II models for each model I coefficient in the same way. Thus, for example, they modeled the log odds of angiography for black versus nonblack persons in each state also as a function of region and angiography availability.

This approach recognizes several sources of variation: (1) Within the same region, for a given level of angiography availability, states vary; (2) within

state, for a given set of patient characteristics, patient outcomes vary; and (3) variation remains after accounting for both patient and state characteristics. Differences in observed rates across states reflect all three sources of variation. This approach is similar to the empirical Bayes method, which recognizes two sources of variation, within units and across units. Thus, empirical Bayes analysis is a special case of hierarchical modeling that produces the same types of shrunken estimates. For example, the log odds of angiography in a particular state is a weighted combination of the intercept from the model that includes only patients from that state (stage I model) and the predicted value from the stage II model based on the region and availability of angiography in the state. The coefficient associated with the effect of race on angiography is a weighted combination of the coefficient from the stage I model and the predicted value from the stage II model. As in empirical Bayes estimation, the degree of shrinkage is a function of the reliability of the within-unit estimate (here, within state) and the estimate of variation across states.

Interval estimates of parameter values from hierarchical models "quantify uncertainty," although they are not CIs.[11] Although terminology varies, Bayesian intervals are generally called *credible intervals*. Goldstein and Spiegelhalter (1996) used credible intervals to examine the New York State CABG mortality report data. They found very wide intervals, which precluded definitive conclusions about most surgeons. For example, the analysis supported strong conclusions about whether rankings fell into the top or bottom half for only 2 of 17 surgeons. Green and Wintfield (1995, 1230) criticized New York's CABG report because "in one year 46 percent of the surgeons had moved from one half of the ranked list to the other." Goldstein and Spiegelhalter (1996, 404) noted that "such variability in rankings appears to be an inevitable consequence of attempting to rank individuals with broadly similar performances." Furthermore,

> An over-interpretation of a set of rankings where there are large uncertainty intervals . . . can lead both to unfairness and to inefficiency and unwarranted conclusions about changes in ranks. In particular, apparent improvements for low ranking institutions may simply be a reflection of "regression to the mean" (Goldstein and Spiegelhalter 1996, 439).

Hierarchical models deal comprehensively and appropriately with the problem of multiple comparisons, as both point and interval estimates for each provider derive from all the data rather than from data specific to that particular provider. Greenland (2000b, 920) notes:

> Giving the target parameters random components [as is done in hierarchical models] treats the problem [of multiple comparisons] with a global loss function quite different from that in classical adjustment . . . modeling of the sort described here attempts to minimize estimation error by using additional background information, while classical methods only attempt to preserve global α-levels through purely arithmetic adjustment.

> It should come as no surprise, then, that critics of the latter find mixed [that is, hierarchical] modeling more acceptable.

In reanalyzing CABG mortality data from the Pennsylvania Health Care Cost Containment Council, Localio and colleagues (1997, 280) used simulations to demonstrate "the dramatic reduction in the number of false outliers with the use of hierarchical statistical models. The hierarchical models maintained adequate statistical power for detecting true departures from expected rates of mortality."

Hierarchical models rapidly become complex, requiring analysts to use computer-intensive simulations to solve them. Software is constantly evolving. The software package BUGS (Bayesian Inference Using Gibbs Sampling) is available free of charge from the United Kingdom's Medical Research Council Biostatistics Unit at the University of Cambridge Institute of Public Health (see www.mrc-bsu.cam.ac.uk/bugs for information on downloading and using the software). BUGS obtains solutions to the models using Markov Chain Monte Carlo (MCMC) simulation methods. This approach is powerful, although some statistical sophistication is required to use it. The HLM software (Raudenbush et al. 2000) is widely used to estimate parameters of hierarchical models with either ML or REML methods. It is more user-friendly than BUGS. Standard statistical packages may also be used to estimate hierarchical models (Singer 1998), although they are not as flexible as specialized software.

One advantage of simulation-based methods is their usefulness for estimating policy-relevant outcomes. Normand, Glickman, and Ryan (1997) used simulation to profile hospitals for Medicare's Cooperative Cardiovascular Project in the early 1990s. Outcomes included the probability that hospital-specific mortality for average patients was at least 50 percent greater than median mortality and the probability that the difference between risk-adjusted mortality (calculated for each hospital using a logistic regression model fit to the hospital's patients) and standardized mortality (predicted mortality based on a model developed from all patients) was large. MCMC simulation methods (e.g., as implemented by BUGS) enable relatively straightforward calculations of such statistics.

Example Using Bayes Estimation

As an example of Bayesian methods, we simulated patient-level cost data and then used two approaches to estimate underlying parameters: shrunken estimates from a hierarchical model and averages calculated directly from the data. We illustrate that Bayes estimates are likely to be more accurate than traditional estimates and show how Bayes intervals and traditional CIs compare.

We assumed that the cost data were generated according to a slight enrichment of the simple hierarchical model described earlier. Specifically, each patient's cost was randomly sampled from a lognormal distribution

with parameters that varied from provider to provider. For each provider, the parameters for the lognormal distribution were randomly sampled from a common normal distribution (truncated to be positive).[12]

We estimated parameters for the underlying distributions on the basis of one year of charge data for patients aged 65 or older admitted to Massachusetts hospitals in Diagnosis-Related Groups 89/90, simple pneumonia and pleurisy. Average costs (more precisely, charges) per patient were roughly $6,400, with an *SD* of approximately $4,000. We assumed hospitals' true mean costs varied from about 20 percent below to 20 percent above the average of $6,400.

For 25 hospitals, we generated data under two alternative assumptions about patient volume, first assuming that each hospital treated 30 patients, often the minimum considered acceptable for producing profiles, and second assuming that each hospital treated 100 patients, a relatively large number for a condition-specific profile. We generated five sets of simulated data for each scenario. From the simulated data, we estimated each hospital's mean two ways: (1) as the average of the data for the hospital and (2) using a hierarchical model with noninformative priors on unknown parameters (i.e., mean and *SD* of the normal distribution and one of the parameters of the lognormal distribution). We ranked hospitals from high to low cost on the basis of their mean costs estimated each of the two ways. We then compared these rankings to rankings based on their actual mean costs, which we knew because we had simulated the data. We reported two measures: (1) on average, how far hospitals' ranks were from their true ranks (calculated as the average of the absolute value of the difference in ranks) and (2) how often hospitals' ranks were five or more ranks away from their true ranks (which would put them in a different quintile).

With sample sizes of 30, ranks based on the mean cost of a hospital's patients were, on average, 4.8 positions away from the true ranks; more than 40 percent of the ranks were five or more positions away from the true ranks. Ranks derived from the hierarchical model were, on average, 2.1 positions away from the true ranks; fewer than 8 percent of the ranks were five or more positions away from the true ranks.

We found similar results with sample sizes of 100, a size usually thought sufficient to obtain good measures. Ranks based on the mean cost of a hospital's 100 cases were, on average, 3.6 positions away from the true ranks; more than 35 percent of the ranks were five or more positions away from the true ranks. Ranks derived from the hierarchical model were, on average, 1.6 positions away from the true ranks; fewer than 5 percent of the ranks were five or more positions away from the true ranks. Simulation results, however, probably overstate the real value of Bayesian estimation because we used the correct underlying model to make the Bayesian estimates. In real life, the hypothesized model is only an approximation of the underlying reality.

Exhibit 12.6 shows, for sample sizes of 30, the raw averages and shrunken estimates for the 25 hospitals from one replication of the simulation. (To avoid cluttering the graph, we connect only some pairs of estimates.) The shrinkage is evident. Estimates for hospitals at the extreme are "pulled" toward the mean, suggesting how much raw averages overestimate differences among hospitals as compared to the estimates from hierarchical models. Because all sample sizes are equal, differences in the amount of shrinkage are caused by differences in the distribution of hospitals' cost data, especially the influence of outliers. Consider the two most expensive hospitals. The most expensive hospital's shrunken estimate was less than that of the second most expensive hospital. The explanation is that two very expensive outlier cases caused the most expensive hospital's high average costs. In contrast, the second most expensive hospital had many cases with relatively high costs but no extreme outliers. The Bayesian estimates discount outliers and shrink the estimates more when outliers drive the raw averages.

Exhibit 12.6 also demonstrates that hierarchical models do not necessarily shrink all estimates toward the overall mean. In fact, the hierarchical model pulled the tenth most expensive hospital's cost away from the mean. The distribution of this hospital's costs is informative. This hospital had many inexpensive cases, but one very costly outlier increased the average. After reducing the effect of this outlier, average cost fell, and the shrunken estimate pulled this "average with outlier effect reduced" toward the mean.

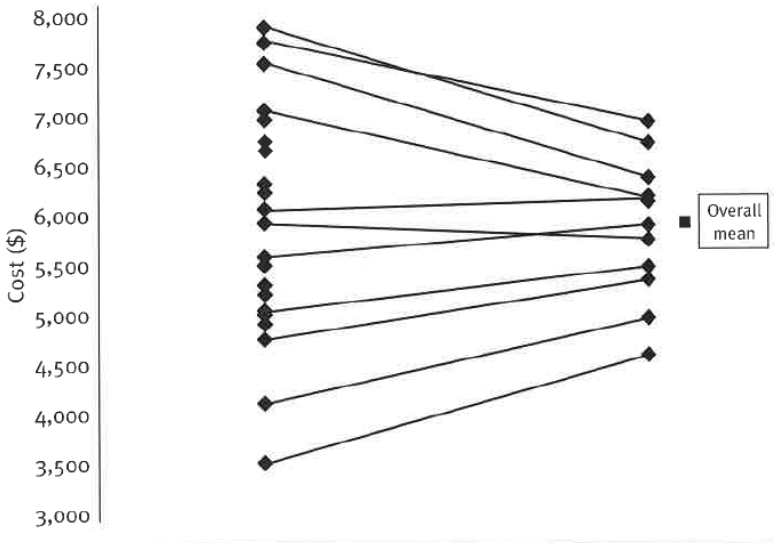


EXHIBIT 12.6 Average Cost (left) and Bayes-Estimated Mean Cost (right) by Provider*

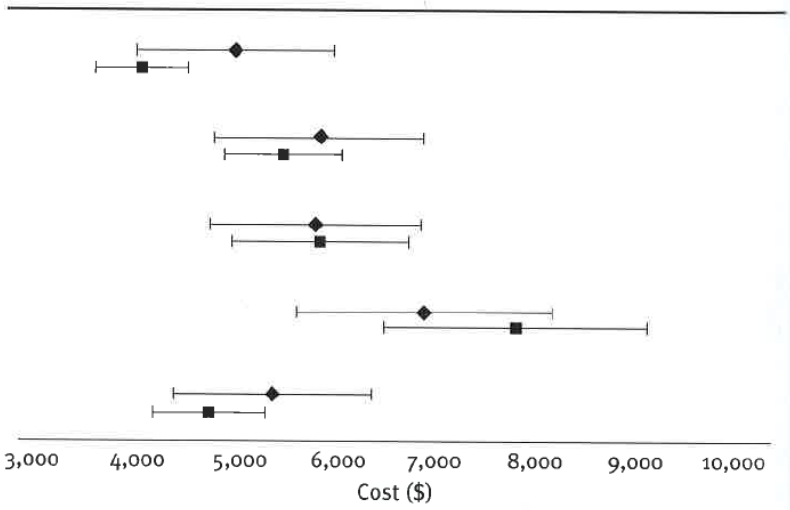
*Calculated from simulated data for 30 patients at 25 hospitals.

For five hospitals, Exhibit 12.7 shows the estimated means as well as CIs and Bayesian credible intervals. Exhibit 12.7 makes two points. First, Bayes estimates pull extreme averages toward the overall sample mean, which is slightly less than $6,000; second, Bayesian credible intervals are frequently wider than CIs.

Hierarchical modeling provides an attractive framework for estimation when one is profiling providers. Shrunken estimates appropriately adjust for the influence of outliers and the increased unreliability associated with estimates from smaller samples. Furthermore, the Bayesian credible intervals better reflect the uncertainty of the estimates than do traditional CIs. Nevertheless, until recently, hierarchical models have generally not been used to profile provider performance outside of research settings. One criticism is the extent to which results are based on the underlying probability models, although as Greenland (2000c, 164) notes:

> Every inferential statistic (such as a *p*-value or confidence interval) is model based, in that some set of constraints (i.e., a model) on the data-generating process must be assumed in order to derive tests and estimates of quantities of interest. . . . Multilevel modeling is distinguished only by its unfamiliarity, which obliges one to make more effort to explain the model. But multilevel modeling need not involve stronger assumptions than ordinary modeling, and in fact provides an opportunity to use weaker assumptions.

EXHIBIT 12.7 95 Percent Bayesian Credible Interval* and 95 Percent Confidence Interval** for Simulated Data for Five Providers



*Top interval of each pair, mean indicated by a diamond.

**Bottom interval of each pair, mean indicated by a square.

In 2007, when the Centers for Medicare & Medicaid Services (CMS) began to make risk-adjusted hospital profiles publicly available on the Hospital Compare website, the use of hierarchical models moved from the research setting to the policy arena. Hierarchical models developed by Krumholz and colleagues (2007) were used to calculate risk-adjusted 30-day mortality rates for two conditions—AMI and heart failure (pneumonia was added in 2008)—at over 4,000 acute care hospitals (see chapters 9 and 17). On the basis of this analysis, each hospital was flagged for each condition on CMS's Hospital Compare website if its risk-adjusted mortality was greater or less than expected for that condition. As noted by Silber and coauthors (2010), the vast majority of hospitals were not flagged; their performance was as expected. The situation for AMI was most extreme: In 2007, 99.5 percent of 4,477 hospitals were shown to have AMI mortality rates not detectably different from the US national rate, 17 hospitals had better rates, and 7 had worse rates; in 2008, of 4,311 hospitals, only 9 had better rates and none had worse than average rates. Are no finer distinctions among hospitals possible? Do these estimators shrink too much, or perhaps not in the best possible way? It is useful to consider these and related questions in the context of a broader literature in statistics and econometrics that distinguishes between random effects (RE) and fixed effects (FE) models.

Random Effects and Fixed Effects Models

Bayesian analyses typically use data from a specific facility (or provider) along with data from a larger group to produce estimates. However, as noted earlier, estimators that stabilize limited information from individual facilities by shrinking their statistics toward a group mean are not necessarily Bayesian. The key distinction for estimating facility effects is between RE models that lead to shrinkage and FE models that do not. By *fixed effect*, we do not necessarily mean that the effect is constant across facilities but rather that each facility's effect is viewed as a fixed value and no data from other facilities contribute information about its performance.

Both RE and FE models traditionally use maximum likelihood algorithms to derive estimates and have long histories in the econometrics literature. Econometricians typically view the effects of local characteristics, such as individual facilities, as nuisance parameters and therefore have had little interest in how the choice of model affects facility-level estimates. Indeed, differences between the econometrics and statistics perspectives largely result from how models are conceptualized and applied. When one is choosing between RE and FE models for provider profiling, the key functional distinction is that the RE choice leads to shrunken estimates and the FE choice to nonshrunken estimates.

Choice of RE Versus FE Model

In basic statistics courses, the need to choose between FE and RE models usually arises when analysis of variance is discussed. Students learn that FE models are appropriate to use when the units studied are the only ones of interest but that they should use RE models when the analytic sample is from a larger population to which findings are to be generalized (e.g., determining whether teaching hospitals or community hospitals treat sicker patients on the basis of data from only a sample of each kind of hospital). However, as Freedman (2005, 86) asserted, any set of units can be considered as "chosen at random by Nature." Hence, the use of RE models can be justified even when the particular units being studied constitute the entire universe of interest. DeLong and coauthors (1997, 2660) provide a pragmatic rationale for using RE models in almost any situation:

> One commonly held belief is that if providers represent the total population about which conclusions are to be drawn, rather than a sample from it, one should consider the provider effects as fixed. A counter argument is that, regardless of the inference, if one can consider provider effects as random realizations from some underlying distribution, then their estimation (or prediction) is more robustly accomplished by treating them as such.

Of course, one **always** can consider provider effects as random realizations of an underlying process. The key issue is to understand the consequences of using RE or FE models for a specific purpose, such as profiling provider performance.

In contrast to traditional statistical teaching, econometricians typically distinguish between FE and RE models when discussing panel data—that is, multiple observations within units (such as states or hospitals) observed at different times. The usual goal is to clarify associations between the outcome and one or more variables of interest, not including facility effects, which are viewed as "nuisance parameters." This goal is achieved by creating a model with coefficients (for variables other than facility effects) that are robust to unobserved effects that may differ across facilities but remain fairly constant within a facility over the time of interest.

For example, Ryan and colleagues (2009) examined the relationship between CMS's process of care measures and mortality for AMI, heart failure, and pneumonia from over 3,000 hospitals in 2004, 2005, and 2006. They used an FE model to control for unobserved hospital-level factors that might confound the relationship between care processes and mortality. For example, a better-organized hospital might have both better process adherence and lower mortality, even if process adherence itself does not save lives. Broadly, the analysis examined the relationship between adherence and mortality over time in each hospital and then aggregated this information across hospitals.

Hanchate and collaborators (2010) used cross-sectional data from 1998–1999 to examine the relationship between surgical volume and mortality following CABG surgery, alternatively using FE and RE models to control for unobserved hospital-level factors. Similar to the Ryan study, their FE analysis examined the relationship between volume and mortality across surgeons in each hospital and then aggregated that information across hospitals. To identify factors affecting patient outcomes, FE models compare outcomes only for patients in the same facility. RE models compare outcomes within and across facilities, producing narrower CIs for coefficients of interest (as long as the coefficients of interest are not facility effects) (Allison 2005).

Why not always use RE models? As noted by Kennedy (1998, 227), "The random effects model has a major drawback . . . it assumes the random error associated with each cross-sectional unit is uncorrelated with other regressors (i.e., predictors or independent variables), something that is not likely to be the case." However, this "drawback" reflects the way in which RE models have traditionally been formulated by econometricians; it is not intrinsic to RE modeling, as we clarify in the next section.

Specifying FE and RE Models

Performance profiles do not control for unit effects; rather, determining the estimated vector of unit effects is the principal objective. For simplicity, assume a single continuous-valued risk factor X_{ij} for patient i at hospital j. Think of X_{ij} as a risk score incorporating many risk factors and "centered" (by replacing X with $X - \overline{X}$) at 0. Thus, $X_{ij} = 0$ denotes average level of risk, positive scores denote riskier-than-average (sicker) patients, and negative scores denote healthier ones.

We assume a dichotomous outcome: $Y = 1$ if the patient dies within 30 days of hospital admission and 0 otherwise. Let $D = Probability\,(Y = 1)$, D_{ij} denoting this probability for the i^{th} person at the j^{th} hospital. As described earlier, a logistic regression model is often used to risk-adjust before profiling:

$$logit(Y_{ij}) = ln(Odds_{ij}) = ln(D_{ij}/(1 - D_{ij})) = b_0 + b_1 X_{ij}.$$

Because we centered the X_{ij}s, b_0 is the natural log of the odds of the outcome for a patient with average risk. Note that this model accounts only for differences in patients' risks. The model ignores hospital effects, so it does not address the possibility that patients at the same hospital are likely to be more similar to each other in terms of important unmeasured risk factors than to patients from different hospitals. In reality, as noted throughout this book, persons distribute nonrandomly across health care providers, including hospitals.

In traditional profiling, we first fit the model to data to identify estimates $\hat{b}_0$ and $\hat{b}_1$ for b_0 and b_1, respectively. We then calculate *PRED* for each individual:

$$PRED_{ij} = \frac{e^{\hat{b}_0 + \hat{b}_1 X_{ij}}}{1 + e^{\hat{b}_0 + \hat{b}_1 X_{ij}}}$$

The sum of the *PRED*s for the individuals in a facility is its expected number of deaths, which we then compare to its observed number of deaths. As described earlier, CIs for each facility's *O/E* ratio are calculated by treating its expected number of deaths as a known constant and its observed number as the result of a binomial or Poisson experiment.

An FE model for profiling appends dummy (indicator) variables for facilities to the logistic regression model provided at the beginning of this section (later we discuss alternatives for coding the indicator variables):

$$logit(D_{ij}) = b_0 + b_1 X_{ij} + \sum_j a_j F_j.$$

As noted, econometricians generally use such models to explore their interest in b_1, not the a_js. "The dummy variable coefficients are inserted merely for the purpose of measuring shifts in the regression line arising from unknown variables" (Kennedy 1998, 227).

Simple RE models may resemble FE models:

$$logit(D_{ij}) = b_0 + b_1 X_{ij} + a_j.$$

The difference is that each a_j is assumed to be a normally distributed random variable (not a fixed but unknown number) with mean 0 and unknown variance $a_j \sim N(0, \tau^2)$. This model is equivalent to the hierarchical model described previously. The key assumption in this formulation is that the a_js are independent of the X_{ij}s. If the X_{ij}s are fixed quantities (i.e., simply facts about the patients and not variables), this statement makes no sense. However, following Freedman (2005), it does make sense if we view the X_{ij}s as having been "selected randomly by Nature."

This independence assumption has important implications. Consider a risk factor that is usually unmeasured and negatively correlated with risk, such as functional status. Patients with higher risk scores tend to have lower (i.e., worse) functional status. Hence, when facility j's average risk score ($X_j = \Sigma_i X_{ij}/n_j$) is greater, its average unmeasured functional status is likely to be lower. In an FE model, the effect of average lower functional status in facility j is reflected in a lower estimate of a_j, which in addition to capturing the effect of interest (i.e., effect of the facility on the outcome) also reflects the effect of the unmeasured confounder correlated with the other predictors. An RE model that does not account for such a confounder may produce biased estimates of both the effect of measured *X*s on the outcome and the facility effects.

The Hausman test for correlation between the random effect and other predictor variables in the model assesses whether the RE model is appropriate. Essentially, it uses a chi-square test statistic to see if there are significant

differences between the two estimated variance-covariance matrixes: one arising when ordinary least squares is used in an FE model and the other when generalized least squares is used in an RE model (Kennedy 1998, 227).

However, though econometricians do not commonly do so, RE models can be formulated so that failure of the independence assumption does not lead to biased estimates of coefficients of individual risk factors or facility effects. Let X_j (i.e., the average value of X at the j^{th} facility) be a set of explanatory variables. Specifically, we model

$$logit(D_{ij}) = b_0 + b_1 X_{ij} + b_2 X_j + a_j,$$

where a_j is assumed to be independent of the other variables and distributed $N(0, \tau^2)$. In this model, b_1 is the difference in outcomes (on a logit scale) between two patients in the same facility whose risk score differs by 1; b_2 is the difference in average outcomes between two facilities whose average risk score differs by 1. b_2 picks up the effect of any unmeasured confounder that differs by facility and is correlated with measured risk factors (e.g., functional status in the previous example). A test of the null hypothesis that $b_2 = 0$ is a version of the Hausman specification test (Snijders and Bosker 1999).

Although the FE model may seem a more straightforward way to account for correlations between the effects and independent variables, it raises its own questions. Substantially less variation is often present in the Xs within each unit (e.g., hospital) than across units. Thus, if the Xs are measured with error, the variation in Xs within units may reflect more "noise than signal." As a result, the coefficient associated with X_{ij} may be biased toward zero (Johnston and DiNardo 1997).

When profiling facilities, we are interested in the coefficients associated with the individual effects (i.e., the a_js). RE models commonly assume the distribution of the effects is normal, though if parameters are estimated using Markov Chain Monte Carlo methods (e.g., as in BUGS), different distributions can be used. FE models do not require *a priori* assumptions about the distribution of the effects.

Coding Fixed Effects

Because econometricians are usually interested in b_1, they pay little attention to how the fixed effects are coded. However, when one is profiling provider performance, coding affects interpretability. Suppose there are J facilities. Commonly, we would code $F_j = 1$ for observations from facility j ($j = 1, \ldots, J-1$) and 0 otherwise. The fit of the model is not affected by which facility is omitted. The model is then fit with the $J - 1$ dummy variables. For the omitted facility, which can be thought of as the reference facility,

$$logit(D_{ij}) = b_0 + b_1 X_{ij},$$

while for each of the other $J - 1$ facilities,

$$logit(D_{ij}) = b_0 + b_1 X_{ij} + a_j.$$

In this formulation, b_0 is the predicted outcome for a patient of average severity at the J^{th} facility, and $b_0 + a_j$ is the predicted outcome for a patient of average severity at the j^{th} facility ($j = 1, \ldots, J-1$). Thus, each a_j estimates the amount by which the effect of facility j differs from the effect of the reference facility, J.

For performance profiling, analysts may want to ask directly which facilities differ from average. For this purpose, we can let the F_js be defined as previous for $j = 1, \ldots, J-1$, but for facility J we let all $J-1$ F_js take the value -1. Thus, for the J^{th} facility,

$$logit(D_{ij}) = b_0 + b_1 X_{ij} - \sum_j a_j,$$

while for each of the other $J-1$ facilities, the logit is as previously presented. With this coding, and with weights (for calculating the average) that count each facility (and its patients) equally regardless of the number of cases contributed, b_0 is the predicted outcome (on a logit scale) for a patient of average risk at a facility with average effect.[13] For patients of similar risk, a_j indicates how much the predicted outcome at facility j ($j = 1, \ldots, J-1$) differs from what is expected at an average-quality facility, and $-\sum_j a_j$ indicates how much the predicted outcome at the J^{th} facility differs from average. Standard errors associated with the estimated coefficients, adjusted to account for clustering (Williams 2000), can be used to identify facilities that are statistically significantly different from average.

Random Versus Fixed Effects Models for Provider Profiling

Models can be used for performance profiling in several ways. As described earlier, classic logistic regression does not include facility effects. Rather, one calculates the expected number of outcomes or events (e.g., deaths) at facility j by summing the individual $PRED_{ij}$s and then computing the ratio of the observed to expected number of outcomes at the facility. We can also extract an expected value for each person, using either an RE or FE model, by applying the model coefficients to the person's characteristics X and setting all the facility effects (the a_js) to zero.

Researchers have explored the extent to which the coefficients associated with the risk factors (Xs) differ as a function of how (or whether) facility effects are included in the model. Austin, Tu, and Alter (2003) used both classic logistic regression and hierarchical logistic regression (equivalent to an RE model) frameworks to estimate models for almost 120,000 patients with AMI admitted to Ontario hospitals. They concluded that "the impact

of patient-level factors on outcomes was virtually equivalent using multilevel and traditional statistical techniques" (Austin, Tu, and Alter 2003, 30). These investigators examined only three risk factors: age, sex, and an illness severity score derived from the Ontario AMI mortality prediction model (Tu et al. 2001). Li and coauthors (2009) compared three approaches—classic regression, FE, and RE models—for developing logistic risk adjustment models to predict decline in activities of daily living among over 600,000 nursing home residents from more than 9,000 nursing homes. Their models included over 20 risk factors. They note "the coefficients (or odds ratios) for resident risk factors estimated by these models were all in the expected direction (i.e., showed construct validity) and in general similar across models (i.e., showed high convergent validity)" (Li et al. 2009, 95).

Both RE and FE models support estimation of CIs or credible intervals for facility effects. Interval widths from FE models reflect only patient-level variation within facilities, while interval widths from RE models reflect variation within, as well as across, facilities. As illustrated in Exhibit 12.7 and discussed earlier, intervals from RE models are usually wider—one reason RE models detect fewer outlier facilities. However, the main reason RE models detect fewer outliers is that their point estimates are shrunk toward a group average.

CMS Hospital Compare RE Model

For a third approach to profiling facilities, Krumholz and colleagues (2007) developed a multilevel RE model for CMS to predict outcomes from risk factors. Using their notation, for patient i discharged from hospital j, we let Y_{ij} be a dichotomous variable indicating whether the patient survives 30 days past admission; also, let $\{Zijk\}$ be a set of risk factors for this patient. They considered the following model:

$$logit(Y_{ij}) = \mu + \omega_j + \sum_k b_k Z_{ijk},$$

where μ is the log of the odds of death for a patient from a hospital with average effect and ω_j is a random effect distributed normally with mean 0 and variance τ^2, $\omega_j \sim N(0, \tau^2)$. ω_j is an indicator of hospital j's performance on a logit scale, and τ^2 is a measure of the variation in performance across hospitals.

The probability that patient i from hospital j dies is

$$\frac{e^{\mu + \omega_j + \Sigma_k b_k Z_{ijk}}}{1 + e^{\mu + \omega_j + \Sigma_k b_k Z_{ijk}}} = \frac{1}{1 + e^{-(\mu + \omega_j + \Sigma_k b_k Z_{ijk})}},$$

while the probability that she would die if at a hospital with average performance is

$$\frac{1}{1 + e^{-(\mu + \Sigma_k b_k Z_{ijk})}},$$

First, one estimates the model parameters μ, the ω_js, and the b_is. Then, one calculates, for each patient, a predicted probability of death, $PRED_{ij}$, based on the patient's risk factors and hospital, and an expected probability of death, E_{ij}, if the patient were treated at a hospital with average performance (i.e., $\omega_j = 0$).

Let $\overline{Y}$ be the overall death rate in the population. Krumholz and collaborators (2007) calculate a risk-standardized mortality measure ($RSMR$) for each hospital as:

$$RSMR_j = \left[\frac{\Sigma_i PRED_{ij}}{\Sigma_i E_{ij}}\right]\overline{Y}.$$

They use a sophisticated bootstrapping approach to estimate CIs for the $RSMR_j$s. This method accounts for uncertainty about the coefficients in the risk adjustment model, the b_is, and the random effects, the ω_js.

In this approach, facility performance is not directly measured by the random effect, and no O/E ratio is calculated. Rather, performance is measured by a ratio in which the denominator is a hospital effect–free expected rate of death (i.e., predicted number of deaths if the hospital's patients were treated at a hospital with average performance) and the numerator is a hospital-specific expected rate of death (i.e., predicted number of deaths based on a best estimate of the hospital's effect). This approach removes the randomness associated with the observed outcome O from the numerator. To date, few studies have evaluated the implications of this approach versus a more standard O/E approach. However, two recent studies have evaluated alternative approaches using both real and simulated data (see discussion in the next section).

In each of 2007 and 2008, CMS used one year of data and the $RSMR$ to profile hospital quality. In 2009, to increase the precision of reported estimates, it started pooling the three most recent years of mortality data for profiling. The 2010 Measure Maintenance Technical Report released by CMS does not give the number of outlier hospitals identified for AMI, heart failure, and pneumonia (Bernheim et al. 2010). Instead, it shows the spread of $RSMR$ values. In the 2010 Measure Maintenance Technical Report (Bernheim et al. 2010) for the years 2006–2008, the minimum, 25th percentile, mean, 75th percentile, and maximum $RSMR$s (expressed as percentages) are as follows: AMI: 10.6, 14.6, 15.8, 16.9, and 24.6; heart failure: 6.1, 9.8, 10.9, 11.9, and 19.4; and pneumonia: 6.8, 10.0, 11.4, 12.5, and 20.7. The odds of mortality for a patient treated at a hospital whose risk was one standard deviation above the national average divided by the odds of mortality for a patient treated at a hospital whose risk was one standard deviation below it was 1.55 for AMI, 1.59 for heart failure, and 1.68 for pneumonia. Thus, although few hospitals were identified as outliers, a fairly large spread of $RSMR$s across hospitals was reported.

Simulation Studies of Profiling Approaches

Researchers have applied various models and profiling methods to real and simulated data and compared the outliers identified by different approaches. With real data, there is no gold standard, and thus we do not know which model is right (Shahian et al. 2010). Simulation allows us to create data sets in which true outliers are known. Thus, we can compare how well different approaches perform in terms of identifying true outliers (sensitivity) and not flagging non-outlier facilities (specificity).

Generation of simulated data is relatively straightforward if analysts assume no performance differences across providers; they can then evaluate models according to the number of providers "falsely" identified as an outlier (e.g., see Localio et al. 1997). However, such an analysis does not address the concern that "true" outliers may be missed—an issue that existed in the original CMS models, which classified almost all hospitals as performing "as expected." Many feel that the current CMS models still do not identify hospitals that may be true outliers. To examine this problem, analysts must build performance outliers into simulated data sets. Unfortunately, innumerable ways exist for outliers to perform differently from expected. Thus, how researchers generate data including true outliers matters.

Using simulated data, Austin, Alter, and Tu (2003) compared using an O/E ratio (which they called an FE model) and an RE model for classifying hospitals as mortality outliers. The data were derived from more than 18,000 AMI patients from 110 hospitals in Ontario. Case mix was adjusted by using the Ontario AMI mortality prediction model (Tu et al. 2001). In one set of simulations, the top 10.5 percent and the bottom 8.5 percent of hospitals were defined as high and low outliers, respectively. In a second set of simulations, a trimodal distribution—a small group of normally distributed low outliers, a large group of normally distributed non-outliers, and a small group of normally distributed high outliers—was assumed. They found that the models' performance depended on hospital size, the mortality distribution, and illness severity. They concluded:

> [T]he sensitivity of the random-effects method was inferior to that of the fixed effects model, whereas under most of the scenarios examined, the specificity of the random-effects model was greater than that of the fixed-effects model. . . . When all hospitals were high-volume hospitals, the sensitivity of the random-effects model improved substantially . . . However, even when all hospitals were treated as high-volume centers, the random-effects model still had lower sensitivity than the fixed-effects model . . . In the unimodal setting, the fixed effects model had substantially higher sensitivity and only marginally lower specificity than the random-effects model (Austin, Alter, and Tu 2003, 536).

Kipnis, Escobar, and Draper (2010) considered FE and RE models, using observations at either the facility level or the individual level. They

applied these models to both real and simulated data. The real data, which came from 17 hospitals in the Northern California Kaiser Permante Medical Care Program, consisted of a risk score X_{ij}, which was a predicted probability of death for patient i at hospital j from a previously developed risk adjustment model, and an indicator for whether the patient died. They fit the following models to these data:

1. FE aggregate-level data: observed mortality divided by expected mortality, where expected mortality equals X_j
2. RE aggregate-level data: predicted mortality divided by expected mortality (the CMS approach) (In this second model, the numerator was $1/(1 + e^{-(a_j + b_j X_{ij})})$ and the denominator was $1/(1 + e^{-b_j X_j})$. Two slightly different versions of this model were considered, one estimated using pseudo-likelihood techniques and the other a Bayesian version, estimated using Markov Chain Monte Carlo simulation.)
3. FE individual-level data: observed mortality rate divided by expected mortality rate, where the expected mortality among the n_j patients in facility j was

$$\frac{\sum_i \frac{1}{1 + e^{-(b_0 + b_j X_{ij})}}}{n_j}$$

4. FE individual-level data: $logit(p_{ij}) = b_0 + b_j X_{ij} + a_j$, with the facility effect estimated as the antilog of a_j
5. RE individual-level data: This formulation was the same as model 4, but the a_js were treated as random effects.

When the real data were used, the FE models identified more outliers than the RE models did (6 to 8 of the 17 hospitals versus 3 of the 17 hospitals, respectively). Despite this difference in performance, the models' hospital rankings were highly correlated (0.91 and higher). This agreement in part reflects the large number of patients in the hospitals considered, meaning there was relatively little shrinkage.

They considered two simulated scenarios: First, each hospital's mortality rate was set equal to its expected rate based on the risk adjustment model, and second, the hospitals' mortality rates were increased, each to a greater degree than the last across the 17 hospitals. When the simulated data including no true outliers were used, the RE models had the highest specificity (99 percent or higher). The aggregate-level FE model also had high specificity (97.5 percent), and the individual-level FE models had lower specificity (87 to 89 percent). In the data with true poor-performing hospitals, both FE and RE models had high sensitivity (mostly higher than 99 percent) when mortality was more than 1.8 percent lower than expected or 2.6 percent higher

than expected. When mortality was within 1 percent of expected, the FE models had relatively high sensitivity (usually higher than 95 percent) but the RE models had poor sensitivity (lower than 20 percent). Kipnis, Escobar, and Draper (2010) concluded that the aggregate-level FE model had the highest specificity and sensitivity.

Racz and Sedransk (2010) also considered several models:

1. The standardized mortality ratio, which multiplies an *O/E* ratio by the statewide mortality rate (They considered this model both with a traditional approach to determining CIs [using Byar's approximation] and with a Bayesian approach.)
2. FE model with dummy variables for each facility
3. RE model for which CIs were estimated using both maximum likelihood and Bayesian methods

Using data from New York State's Cardiac Surgery Reporting System (CSRS) for the years 1992–1999, they found that the *O/E* model (#1) identified the same outliers whether its CIs were determined using Byar's approximation or a Bayesian approach. The RE model (#3) using both methods of calculating CIs identified 11 out of 255 hospitals as outliers; an additional 9 hospitals, not identified when Bayesian methods were used, were identified when the maximum likelihood method was used. However, in general, for the RE model, the Bayesian and maximum likelihood approaches produced similar intervals. The *O/E* ratio approach (#1) identified 14 hospitals as high outliers; 8 of these hospitals were identified as such by the maximum likelihood RE model. Of 14 hospitals identified as low outliers by the *O/E* ratio approach, 12 were identified as low outliers by the maximum likelihood RE model. Among low-volume hospitals (those having fewer than 200 cases), none of the six hospitals identified as a high outlier by the FE model was identified by the maximum likelihood RE model. Among larger hospitals (those having more than 200 cases), only 7 of the 18 hospitals identified as high outliers by the FE model were identified as high outliers by the maximum likelihood RE model. Racz and Sedransk (2010) generated simulated data sets in which two hospitals were high outliers. For these hospitals, the log odds of mortality were specified amounts (25 percent, 50 percent, and 75 percent) higher than average for the average patient. Analyses of a simulated data set confirmed the findings from the CSRS data: (1) For low-volume hospitals, the RE model was less likely than the *O/E* model to identify the outlier hospitals; (2) when patient volume was high, both the RE and *O/E* models detected the outlier hospitals; and (3) neither method was likely to declare a non-outlier hospital an outlier. (Because hospitals with no deaths must be excluded from an FE model, and because such situations arise with simulated data, results from the FE model were not calculated for the simulated data.)

Racz and Sedransk (2010) described several hospitals that responded to poor outcomes on the New York State reports by launching improvement projects. For example, Bellevue Hospital was declared a high outlier in both 1994 and 1998. It suspended its cardiovascular surgery service and reopened a redesigned service in 2001. Had RE models been used, Bellevue Hospital would not have been declared a high outlier in either 1994 or 1998. Racz and Sedransk (2010) also discussed the case of the University of Massachusetts Memorial Hospital in Worcester, which had not been declared an outlier in the Massachusetts Department of Public Health (DPH) report in 2002. Shahian and colleagues (2005) discussed the fact that this hospital (referred to as "hospital 13") would have been declared an outlier had New York State's methodology been used but was not considered an outlier by hierarchical models. Referring to such discrepancies, they pointed out that the New York State method was more likely to detect false outliers. Shortly before the University of Massachusetts Memorial Hospital was declared an outlier in the Massachusetts DPH report of 2003, the hospital suspended its program. After making substantial changes, the program reopened. In fiscal year 2008, as reported on the CMS Hospital Compare website, survival rates for heart attack patients in the reopened program were better than the US national average.

Silber and colleagues (2010) studied alternative formulations of RE models to address a core problem in profiling: how to report on hospitals with so few cases that their data are inadequate to produce meaningful measures. Technically, their research paper explored the exchangeability assumption underlying RE models. Consider an outcome such as death within 30 days following hospital admission. As noted earlier in this chapter, exchangeability asserts that in the absence of adequate information on 30-day mortality in a particular hospital, analysts can reasonably assume that its mortality rate is average. However, what if low-volume hospitals have systematically worse mortality than that of higher-volume hospitals? In that case, should the shrunken rate of a low-volume hospital be calculated by moving the observed rate toward the average rate of all hospitals? Or should it be calculated by moving the observed rate toward the average rate of other small hospitals? Note that concerns about shrinkage to a possibly inappropriate global average apply whenever a hospital characteristic could systematically affect quality (e.g., public or private ownership, safety-net hospital). However, the concern is greatest when the characteristic is volume because estimates for low-volume hospitals are the most drastically affected by shrinkage.

Silber and colleagues (2010) examined this question in regard to CMS's Hospital Compare RE model (Krumholz et al. 2007) used to profile hospitals on the basis of 30-day mortality rates following admission for AMI. Evidence suggests that AMI mortality is higher in low-volume hospitals. Silber and coauthors (2010) assessed the effect of shrinking the rates of low-volume hospitals

toward the average of other low-volume hospitals versus shrinking toward the average of all hospitals. When volume is small, little weight is placed on the facility-specific mortality rate; most of the weight is placed on the overall average. Thus, low-volume hospitals mostly are reported as average. If hospital volume is used as a facility-level variable in an RE model, the shrunken rate is calculated by moving the observed rate toward a mean for hospitals with similar volumes. In this case, predictions by volume category from the RE model become similar to traditional *O/E* ratios by volume category.

As Silber and collaborators (2010) noted, the modified (volume-included) RE model enables one to determine whether a particular small hospital is different from its peer group of other low-volume hospitals, and they highlighted the fact that small hospitals as a group have higher mortality. However, they also observed that there is a cost: The mortality rate of a high-quality small hospital will be lowered toward the average of all small hospitals, making it appear as a lower performer than it might actually be. They concluded that a "random effects model of whatever kind should not be used to assign grades to the performance of individual providers when their volume is low" (1163). Fundamentally, hospital quality cannot be reliably measured from the experience of very few cases. For this reason, we recommend prominently reporting the number of cases on which a performance profile is based.

Shrinkage Versus Nonshrinkage Estimates

Austin, Alter, and Tu (2003, 537) articulated the policy implications of choosing between FE and RE models:

> [O]ur findings highlight the classic tradeoff between sensitivity and specificity. Random-effects models were more specific than sensitive. Conversely, fixed-effects models were more sensitive, but at the expense of specificity. Which method is best? The answer depends on the goals of policy makers, researchers, clinicians and patients. If the primary objective of publishing hospital report cards is to identify those institutions with higher than acceptable mortality rates, at the risk of falsely labeling a hospital as a poor performer, then fixed-effects models should be employed . . . However, advocates of this approach must bear in mind that fixed-effects models, with their lower specificity, will result in a greater number of false positive[s] . . . Conversely, a more conservative approach may employ random-effects models (with the caveat that a large proportion of hospitals with unacceptably high mortality rates may be classified as being no different than average).

In using the modeling approach of Krumholz and colleagues (2007), CMS has chosen to err on the side of not identifying some true outliers so as to prevent labeling other hospitals as outliers inappropriately.

Krumholz and coauthors (2006a, 2006b), in suggesting standards for statistical models used for public reporting of health outcomes, summarized

many of the previously mentioned advantages of hierarchical models for profiling. Silber and colleagues (2010) cautioned that, even if hierarchical models are used, they can be formulated in various ways and find substantial differences as a result. For example, if low-volume hospitals' performance is moved toward a low-volume hospital mean value, the point estimate of all low-volume hospitals would reflect poorer-than-average performance. However, the interval estimates for these hospitals likely would be very wide, and many of them may not be flagged as outliers. In contrast, if their point estimate is moved toward a grand mean value, all could be viewed as performing "as expected." Other characteristics (e.g., teaching status, ownership or presence of a cardiac catheterization lab) can also be used to shrink the performance of certain kinds of hospitals to different means.

In addition to existing simulation studies (Austin, Alter, and Tu 2003; Kipnis, Escobar, and Draper 2010; Racz and Sedransk 2010), more research is needed to better understand how different models and methods for profiling perform with real and realistic health data. In the more recent simulation studies, the old-fashioned *O/E* ratio, despite its lack of conceptual attractiveness, appears to strike a reasonable balance between sensitivity and specificity.

Outlier identification and the corresponding measures of sensitivity and specificity derive from classical statistical thresholds—the 95 percent CI or the 0.05 alpha value used to evaluate null hypotheses. As noted by Lehrer (2010, 55) in a thought-provoking article, "This ubiquitous [hypothesis] test was invented in 1922 by the English mathematician Ronald Fisher, who picked five percent as the boundary line, somewhat arbitrarily, because it made pencil and slide-rule calculations easier." No particular reason compels adherence to these thresholds. Given any set of facility rankings, one can flag any number of facilities just by shifting the outlier threshold. Perhaps the key to comparing two hospital ranking methods is seeing whether they identify the same hospitals when each is used to flag, say, the 100 most problematic hospitals. If two methods flag different hospitals, it is critical to find ways to decide which approach produces the more credible and useful list.

Comparing Outcomes Over Time

This chapter has concentrated on cross-sectional comparisons—information relating to a single period. However, a powerful profiling tool to use is *longitudinal analysis*, which involves examining changes over time. As Berwick (1996, 4), a leading health care quality improvement expert and head of CMS (July 2010 to December 2011) said, "Pick a measurement you care about, and begin to plot it regularly over time. Much good will follow."

The plotting of measures over time highlights change. For simplicity, suppose all patients have the same level of risk. Suppose we examine the

problem rate in two hospitals in year 1. Hospital A has a problem rate of 3 percent, with a 95 percent CI from 2 percent to 4 percent, whereas hospital B's rate is 5 percent, with a 95 percent CI from 4 percent to 6 percent. From the classical hypothesis-testing perspective, we can reject the hypothesis that the underlying problem rates at the two hospitals (P_A and P_B) were identical in year 1 in favor of the alternative that hospital B's rate was higher. This conclusion does not mean that hospital B's problem rate will be higher than A's next year. Furthermore, even if in year 2 hospital A's problem rate is statistically significantly higher than B's, the assessment of which facility did better in year 1 was not necessarily wrong. Hospital B might have improved.

However, provider profiles are useful chiefly to the extent that they reflect a persistent reality. Green and Wintfeld (1995, 1230) criticized New York State's CABG mortality report, noting:

> The usefulness of the risk-adjusted data was also limited in that surgeons' rankings during two years of the study offered few clues about their position in the subsequent year (R^2 = 4.9 percent). . . . The fact that surgeons' performance ratings fluctuate so much from year to year means that by the time the data are published, users of the report can have little confidence that the ratings are still applicable.

They thus speculate that real differences in comparative performance are outdated by the time they become publicly available. However, Goldstein and Spiegelhalter (1996) provide a more fundamental critique, suggesting that such large changes in rank are likely when true differences in provider performance, even if real and stable over time, are small compared to random "noise."

For the National Surgical Quality Improvement Program (see Chapter 8), Khuri and colleagues (1998) illustrated a way to portray *O/E* ratios over time. Such presentations suggest that providers' performance varies from year to year. From the numbers alone, however, we cannot know how much variability results when some providers improve more than others (i.e., last year's data are outdated) and how much is due to randomness (i.e., noise overwhelms the "signal"). In-depth study is required to disentangle these different possibilities. Nevertheless, longitudinal plots can reveal when providers' ranks change dramatically from one year to the next. When large yearly changes are common, both public reporting and decision making should be restrained. In particular, report cards should not list providers in rank order of their measured performance because this practice reinforces the impression that the figures are reliable. Managers should think twice before disrupting provider–patient and network–provider relationships over findings that may be transitory, even if real.

Given the various methodological concerns and resulting questions about interpretation, profiles should be employed only where they are likely to be useful. For example, if last year's findings differ from this year's

because relative quality changed rapidly among providers, profiling data are most valuable for quality improvement and less useful to large purchasers and individual consumers of health care services aiming to select providers for their future care. Even when longitudinal analyses show stability over time, legitimate concerns remain over whether consistently poor performers are the victims of inadequate risk adjustment. Future research must identify profiling information that is relatively stable over time and distinguish it from figures that fluctuate without obvious explanation. In most current profiling initiatives, random noise is a major consideration, as are unmeasured differences in patient risk. Small sample sizes for individual providers also raise concerns. These factors limit the inferences that can be drawn from practice profiles. Longitudinal plots often are a sobering reality check to the evaluation process.

As described in Chapter 17, despite methodological concerns, profiles are increasingly generated and used as important tools in ensuring health care value, a melding of cost and quality. Comparison of patient outcomes across providers can be valuable, but given the state of the art, reliance on profiles alone to make all-or-nothing business decisions (e.g., withdrawing business from outlier providers) is inappropriate. In this context, profiles are likely to generate (often well-founded) criticism and heighten adversarial relationships among providers, payers, and policymakers. Similarly, if such profiles are disseminated to a public unaware of the need for cautious interpretation, further controversy may erupt, impeding opportunities for useful dialogue and improvements. If providers are given profiles but not educated on how to use them productively to identify areas for improvement, they likely will ignore the information.

Profiles comparing patient outcomes are most valuable in an environment of cooperation and collaboration that offers incentives for learning and improvement. As competitive pressures increase, however, this ideal environment may be more pipe dream than tangible reality.

Notes

1. A useful summary statistic for a nonnegative variable Y is the observed coefficient of variation: $CV = s/\bar{Y}$. When one looks at total costs next year for a heterogeneous population (with many zeros and a few extreme outliers), the CV is often as large as 3 or 4. However, when one looks at costs for patients with a specific condition, the CV can be much smaller.
2. Recognizing that averages based on moderate sample sizes are only approximately normal, we avoid the appearance of precision implied by using intervals with half-widths equal to $1.96SE$. Thus, we use $2SE$.

3. s is calculated as the square root of a weighted average of the SD^2s at each hospital, as described earlier. If the data were distributed normally, Y_A would be in this interval about 95 percent of the time. With highly skewed cost data and modest sample sizes, the probability of it being in this interval is lower. As a result, under the hypothesis of no differences among providers, having more than 5 percent of "normative" providers fall outside these bounds is not surprising.
4. The calculations in Exhibit 12.4 are approximate. Especially when D is near zero, a reasonable 95 percent CI would not be centered at D. For example, after observing five deaths among 500 patients (1 percent), the modified Wald method that Agresti and Coull (1998) recommend yields the 95 percent CI 0.4 percent to 2.4 percent. Note, however, that the half-width of this interval is 1 percent, as suggested by the calculations in the table.
5. A better CI of approximately the same width would be 5.4 percent to 17.6 percent.
6. Odds—the ratio of the probability of an event to the probability that it will not happen—are frequently used in bets. For example, an event has 3-to-1 odds (i.e., odds = 3) when it has a 0.75 chance of happening and a 0.25 chance of not happening.
7. This expected rate is identical to that from a model that uses the single predictor of high versus low risk to predict probability of death from the whole population.
8. Greenland (2000c) gives an excellent nontechnical description of the principles of multilevel modeling. McNeil, Pedersen, and Gatsonis (1992) also provide a nontechnical description of hierarchical models in the context of provider profiling. In a review of cardiac surgery report cards, Shahian and colleagues (2001) discuss problems with traditional approaches and the advantages of using hierarchical models. Normand, Glickman, and Gatsonis (1997) provide a technical discussion of statistical methods for profiling providers.
9. For ease of exposition, we ignore the interesting observation that "you can load a die, but you can't bias a coin" (Gelman and Nolan 2002, 308).
10. Traditional thinking held that statistical estimators should be unbiased (i.e., the difference between the expectation of the estimator and the parameter being estimated should equal zero). However, overall error, measured by the mean square error (MSE), has two components: (1) bias in the estimated parameter and (2) the spread of individual data points around the estimated parameter. Stein estimates are biased, but they have smaller MSEs than the traditional estimate has.
11. A 95 percent CI is an interval that has a 95 percent chance of covering the parameter of interest; if data were resampled from the

population 100 times and 95 percent CIs were constructed, about 95 of these 100 intervals would include the underlying parameter. In this interpretation, the intervals have a chance of covering a fixed parameter. Ninety-five percent CIs are often interpreted incorrectly as being the interval in which the parameter value has a 95 percent chance of falling. In this incorrect interpretation, the parameter value is viewed as having a chance of falling into a fixed interval. The distinction is between the chance that the interval covers the parameter (which is what a CI is) and the chance that the parameter falls in the interval (the way in which a CI is incorrectly interpreted). A non-Bayesian framework assumes that the parameter value is fixed. Hence, considering the chance that the parameter value falls in an interval makes little sense. A Bayesian framework creates a probability distribution for parameters; therefore, considering the chance that the parameter lies in some interval makes sense. "Some interval" is the credible interval determined in a Bayesian analysis.

12. A lognormal distribution with parameters μ_j and σ^2 has mean $e^{\mu_j + \frac{\sigma^2}{2}}$ and variance $(e^{\sigma^2} - 1)e^{2\mu_j + \sigma^2}$. Here, we supposed σ^2 was fixed and that for each provider, the parameter μ_j was generated according to a common normal distribution. Then, we generated patients' costs according to a lognormal distribution with parameters μ_j and σ^2 and calculated the true mean for a provider as $e^{\mu_j + \frac{\sigma^2}{2}}$.
13. For an illustration of this concept in a simple situation, assume we have three facilities. Facility 1 has three patients, facility 2 has four patients, and facility 3 has two patients. Let Y_{ij} be the outcome for patient i in facility j. Assume the following model: $Y_{ij} = a + b_1F_1 + b_2F_2$. F_1 is coded 1 for facility 1, 0 for facility 2, and −1 for facility 3, and F_2 is coded 0 for facility 1, 1 for facility 2, and −1 for facility 3. The expected outcomes are as follows: $Y_{i1} = a + b_1$ for the three patients in facility 1, $Y_{i2} = a + b_2$ for the four patients in facility 2, and $Y_{i3} = a - b_1 - b_2$ for the two patients in facility 3. Let Y_j be the mean outcome in facility j. Then, $Y_1 = (3a + 3b_1)/3$, $Y_2 = (4a + 4b_2)/4$, and $Y_3 = (2a - 2b_1 - 2b_2)/2$. The sum of the three means, $Y_1 + Y_2 + Y_3$, equals $12a + 12b_1 + 12a + 12b_2 + 12a - 12b_1 - 12b_2 = 36a/12 = 3a$. Therefore, the unweighted mean of the outcomes in the three facilities, $(Y_1 + Y_2 + Y_3)/3$, is a.

CHAPTER

13

RISK ADJUSTMENT IN PEDIATRIC POPULATIONS

Karen Kuhlthau, Jeanne Van Cleave, and Timothy G. Ferris

Just as broader societal policies include provisions specific to childhood, risk adjustment for pediatric populations recognizes unique issues pertaining to children. Several factors and characteristics make children special, including relatively low mortality and morbidity rates; the importance of developmental milestones; rapid growth, especially in the first years of life; the need for proxies, typically parents, to represent patients' views and experiences until the age of consent; specialized pediatric services, such as neonatal intensive care; the use of clinicians and institutions outside the traditional acute care delivery system, such as school-based and public health clinics; and the need for adult assistance (Forrest, Simpson, and Clancy 1997). Furthermore, certain important outcomes of interest differ in size (e.g., typical expenditures) or content between adults and children.

Risk adjustment for pediatric outcomes has received relatively less scrutiny than risk adjustment for adult outcomes. This relative lack of attention largely reflects the historical impetus for risk adjustment research. Most early studies of risk adjustment addressed Medicare concerns, such as designing prospective payment systems and examining Medicare hospital mortality rates. Today, interest in pediatric risk adjustment is growing as analysts increasingly grapple with designing pediatric quality measures and setting capitated payments for Medicaid programs, which cover many children nationwide. Some risk adjustment methods initially oriented toward adults now also apply to children, while others (especially those that are disease-specific or designed for specific care locations, such as the neonatal intensive care unit [NICU]) were developed specifically for pediatric populations.

This chapter describes conceptual issues raised by pediatric risk adjustment and reviews current methods. Pediatric risk adjustment is evolving quickly in response to the national impetus to develop quality metrics, so new methods may be available in the near future.

What Is Special About Children?

Pediatricians generally define the pediatric population as newborns up to young adults (see the American Academy of Pediatrics website at www.aap.org).

Infancy includes children aged 0 to 11 months old; the first 28 days of infancy is called the *neonatal period*. *Childhood* is broken down into early childhood (ages 1 to 4 years), middle childhood (ages 5 to 10 years), and adolescence (ages 11 to 21 years). Some consider adolescents over age 18 to be adults, depending on the context for making this distinction. The National Institutes of Health (NIH 1998) defines children as individuals under 21 years of age and requires the inclusion of children "in all human subjects research, conducted or supported by the NIH, unless there are scientific and ethical reasons not to include them." In support of this policy, the NIH (1998) noted that "Most research on the cause, treatment and cure of diseases which affect children rely primarily on adults as subjects in clinical trials. Consequently, treatment options which may be effective for adults can have an adverse impact on the outcome of children as well as on their future growth and development."

The ambiguity over the end of the pediatric period is driven partly by adolescents themselves (preferring to switch to adult-care clinicians), by legal status changes, and by program eligibility. The disease epidemiology and care-seeking behavior of adolescents may be more similar to those of young adults (e.g., reproductive health, pregnancy, acute injury) than to those of younger children (Schmidt and White 2002). For purposes of risk adjustment, arguments can be made for combining adolescents and young adults with adults (e.g., for measures related to pregnancy), treating them as children (e.g., for measures related to pediatric intensive care unit [PICU] outcomes), or categorizing them separately (e.g., measurements related to adolescent screening for sexually transmitted diseases). Among other factors, these decisions depend on the outcome of interest, the nature of potential risk factors, whether youth are treated by adult or pediatric providers, and the characteristics of the patient populations.

Several factors make risk adjustment for children different from approaches designed for adults. The first is the epidemiology of childhood disease and chronic conditions. Recent trends show an increase in the prevalence of childhood chronic conditions (Perrin, Bloom, and Gortmaker 2007; Van Cleave, Gortmaker, and Perrin 2010). Despite this increase, relatively few children have serious persisting illnesses or congenital conditions requiring medical attention. Unlike adults, children rarely have multiple coexisting diseases (van den Akker et al. 1998). Morbidity in childhood typically involves brief, acute viral or bacterial illnesses or accidental injuries. Children generally have much lower average health care expenditures than do adults (Thorpe and Machlin 2001; Machlin 2009; Machlin and Kress 2009). Because of these relatively lower costs, examining health care for children may appear less important than examining health care for adults for purposes of quality improvement, determination of fair payment, or cost containment (Richardson, Tarnow-Mordi, and Lee 1999).

Given this epidemiology, the number of children with a serious health condition is generally small, except in the context of some specialty clinics. Even for relatively common chronic conditions such as asthma, the number of patients may be too small to produce rigorous risk-adjusted outcomes for children treated by individual physicians. Clinicians who treat persons of all ages, such as family physicians, rarely see enough children to assess outcomes for their pediatric patients. In addition, the relatively low use of hospital care by children (Krauss, Machlin, and Kass 1999; Freed, Nahra, and Wheeler 2004) suggests that hospital-based outcome measures are less useful for pediatric populations than for adult populations. Nonetheless, comparisons of outcomes across hospitals or hospital units treating many children (e.g., children's hospitals, NICUs) are often sought for various purposes, such as quality assessment, and these comparisons often require risk adjustment.

Second, important risk factors for children and adults may differ. While diagnoses are often critical risk factors for older adults, developmental milestones, functional status, family supports, and social environment may be more important risk factors for children (Larson et al. 2008; Stein, Siegle, and Bauman 2010). ICD-9-CM, the diagnosis coding nomenclature currently used in administrative databases (see Chapter 5), lacks codes for many developmental indicators, functional status, and familial attributes. Codes for symptoms and signs give little indication of their severity or chronicity. In addition, the clinical meaning of codes may differ between children and adults (e.g., heart murmur, urinary tract infection, chest pain, sleep disturbance, shortness of breath). Although codes do exist for some developmental conditions, they rarely are used because such conditions usually do not require medical interventions. For example, the absence of claims for specific services relating to intellectual disabilities or developmental delays hampers the use of administrative databases to identify children with these conditions, although relevant codes exist[1] (Perrin et al. 1998).

The quality of diagnosis coding in pediatric settings is unknown. Most studies of coding completeness and accuracy have targeted Medicare populations (e.g., Cooper et al. 2000; Lamont et al. 2002), as have regulations, auditing, and oversight of coding and medical record documentation. Diagnosis coding and medical record documentation may therefore be even less accurate and clinically meaningful for children than for adults. As for adults, some conditions in children, such as HIV infection and mental health disorders, are undercoded, perhaps because of fears of "labeling." Attachment of socially disadvantageous diagnostic labels may carry longer-lasting implications for children than for adults. On the other hand, parents may need diagnostic labels for their children to obtain services for them, such as special education or other programs targeting dyslexia or attention deficit hyperactivity disorder.

Third, processes of care sometimes vary between adults and children and thus affect some risk adjustment approaches. For example, some risk adjustment tools rely on pharmacy claims (Kuhlthau et al. 2005), and for many conditions, patterns of medication use differ between adults and children (Ferris et al. 1998; Stafford et al. 1999). Among the ten most frequently used medications for children, only two reliably indicate the presence of a chronic condition, whereas nine of the most common medications for adults are usually prescribed for chronic conditions. Moreover, many drugs are not tested on children, so less evidence is available to guide pediatric medication use.

Fourth, most health care encounter databases include information on services covered by only one public or private insurer, and children often use services from diverse settings inside and outside the traditional acute health care delivery system. Administrative databases generally do not include information on services provided by school-based or public health clinics; similarly, medical records in private doctors' offices typically do not contain information from school or public health settings. Thus, children who use early intervention, public health, and school-based clinics may appear healthier than they actually are because existing data sets do not include complete health-related information. A parallel concern is that some children use both specialized pediatric care systems (e.g., pediatric subspecialists, general pediatricians, and children's hospitals) and general health care delivery systems. Because pediatric care systems likely serve children with more severe conditions (Kuhlthau et al. 2001), risk adjustment may be particularly important for comparing outcomes for children across pediatric and general care systems.

Finally, parental, familial, and community factors, such as the local educational and health care environments, strongly influence children's health and use of health care services. For example, some communities rely heavily on public health authorities to provide pediatric preventive care, while others visit physicians' offices, and not all cities and towns have access to children's hospitals. These complex, interrelated psychosocial and environmental factors may be important risk factors for certain outcomes. For instance, timely delivery of childhood vaccinations is a common quality indicator for children. Although virtually all children require immunization, family characteristics (e.g., single parent, cash assistance eligibility, presence of an additional sibling) are associated with immunization rates (Wood et al. 1995; Alessandrini et al. 2001). Stratification of children by these factors may therefore be important for comparing immunization performance across health plans or providers. However, few databases contain information on these important familial attributes.[2]

The prominent role of the family and the environment in which children live—social and physical determinants of health—highlights the concerns about medical care equity raised by the Institute of Medicine's (IOM 2001a) report *Crossing the Quality Chasm*. Both social and biomedical risks

predict poorer health in children (Stein, Siegle, and Bauman 2010). Ignoring these critical psychosocial, socioeconomic, and community factors could bias comparisons across health plans or providers in performance profiles. Unadjusted comparisons could appear fundamentally unfair to providers who see disproportionate numbers of vulnerable children. As a result, health plans or providers might be leery of enrolling or treating disadvantaged children, thus impeding their access to care.

Risk-Adjusting Pediatric Costs

This section introduces approaches to risk-adjusting cost outcomes for children. As noted earlier, some risk adjustment methods oriented initially or primarily toward adults have incorporated approaches to the pediatric population, while other methods have been developed explicitly for children. The first category of risk adjusters typically focuses on cost outcomes, while the second emphasizes quality indicators (e.g., NICU or congenital heart surgery mortality).

Various risk adjustment methodologies that use administrative data (see Chapter 5) aim to predict future costs (typically total charges or payments) for defined populations, potentially to assist in setting capitated payment levels. Examples include the Adjusted Clinical Groups (ACGs), Diagnostic Cost Groups (DCGs), Chronic Illness and Disability Payment System (CDPS), and Clinical Risk Groups (CRGs) (Muldoon, Neff, and Gay 1997). Pediatricians helped create these methods' clinical algorithms for pediatric populations by grouping together clinically homogeneous ICD-9-CM diagnosis codes relevant to children. Some specifically tested their algorithms on child populations using Medicaid or private insurance databases. (Shenkman and Breiner [2001] provide a detailed tabulation of the characteristics of several of these systems, although this information may be out of date.)

One strength of these methodologies for adults is their ability to quantify the combined effect of multiple coexisting conditions. However, because children typically do not have extensive comorbid illnesses (van den Akker et al. 1998), this advantage may not extend to predicting pediatric costs. Indeed, using various risk adjusters (ACGs, Ambulatory Diagnostic Groups [ADGs], CDPS, and hierarchical condition categories [HCCs] from the DCGs) and Medicaid data from several states, Kronick and colleagues (2000) found lower explained variance for child populations than for adults.

Other studies demonstrate that these risk adjustment methodologies can predict pediatric costs, but systematic errors may occur for different subgroups of children. Using data representing Medicaid-insured and privately insured children, Fowler and Anderson (1996) tested DCGs, ACGs, and Payment Amounts for Capitated Systems (PACS). Each risk adjuster improved the model fit as measured by R^2 statistics. (Improvements were relative to a

model that included only sociodemographic variables.) However, two of the methods underpredicted expenditures for children with chronic conditions and overpredicted costs for children without chronic conditions. Using Medicaid data, researchers examined the predictive accuracy of several risk adjustment models (ACGs, ADGs, HCCs, and CDPS) for simulated groups having different expenditure, hospital, and chronic health characteristics. In general, as expenditures, number of admissions, and burden of chronic conditions in simulated populations increased, the models' predictive accuracy fell below 1, indicating potential underpayment (Hwang, Ireys, and Anderson 2001). A study examining the performance of different risk adjustors (CDPS, DCGs/HCCs, ADGs, ACGs, MedicaidRx, and RxGroups) for children insured by Medicaid showed that the different risk adjusters performed in a similar fashion and that they tended to underpredict expenditures for children in foster care and Supplemental Security Income programs but slightly overpredict expenditures for other Medicaid-enrolled children (Kuhlthau et al. 2005). When the ACGs, DCGs, Global Risk-Adjustment Model (GRAM), prior-expense model, and RxRisk model were compared for the subgroup of children under age 13, the estimates of the area under the receiver operating curve (see Chapter 10) for predicting high-cost cases were similar for all models (Meenan et al. 2003).

The Chronic Disease Status (CDS) method developed at Group Health Cooperative of Puget Sound used pharmacy claims to score individuals on the basis of their chronic diseases, the complexity of their medication regimen, whether the underlying condition was potentially life-threatening, and whether the medications treated diseases and not symptoms. Controlling for health care utilization, the CDS correlated with physician-rated disease severity, patient-rated health, hospitalizations, and mortality (Von Korff, Wagner, and Saunders 1992). The CDS excluded medications that treat diseases of children. Fishman and Shay (1999) built on the CDS to create a pediatric-specific measure. A medication-based risk adjuster, Medicaid Rx, was developed through pharmacologic review and empirical evaluation (Medicaid covers 25–30 percent of all children in the United States). The abilities of Medicaid Rx and CDPS to predict expenditures were compared, and they were found to perform similarly for Temporary Aid to Needy Families (TANF) populations (TANF populations include high proportions of children). The combined model provided the best predictive ability (Gilmer et al. 2001).

Mental health problems are among the most prevalent and morbid conditions in children, particularly adolescents. Mental health and substance abuse services have often been "carved out" of existing health insurance contracts (see Chapter 14). In addition, insurers frequently place caps on the numbers of visits or services for these conditions. Therefore, insurers must pay special attention when setting capitation rates for mental health and substance abuse services.

Using Michigan Medicaid data, one study predicted mental health and substance abuse expenditures for adults and children (stratified analysis) on the basis of demographic characteristics, indicators of psychiatric disability, and existing risk adjusters (ADGs and HCCs). The best model for children included a risk adjuster and the psychiatric indicators; mean prediction errors were smaller for children than for adults (Ettner et al. 2001). The study results were similar when data on a privately insured population (adults and children, stratified by age group) were used and ACGs, ADGs, and HCCs were examined (Ettner et al. 1998). The Mental Health Parity and Addiction Equity Act of 2008—a federal law requiring Medicaid managed care plans and large employer group health plans that offer mental health and substance use disorder coverage to provide those benefits at levels equal to the plans' medical/surgical coverage levels—may change some of these relationships (CMS 2011b).

Risk-Adjusting Pediatric Quality Outcomes

Researchers have predicted pediatric outcomes relating to processes of care, including subspecialist processes (Kuhlthau et al. 2001; Forrest and Reid 2001) and hospitalization and emergency department processes (Perrin et al. 2002). In each of these studies, risk adjustment substantially changed the results. Admittedly, these outcomes are closely correlated with expenditures. For investigations of quality for specific conditions, such as asthma care, risk adjusters designed to consider all diagnoses may not be as meaningful as those created for the specific condition (Shields et al. 2002). Several risk adjusters have been developed for specific pediatric care settings and for particular conditions.

NICU Outcomes

Morbidity and mortality for newborns, particularly those treated in NICUs, are important child outcomes. Not surprisingly, therefore, risk adjustment for NICU populations has received considerable attention. The primary methods of risk adjustment for neonates have undergone extensive testing and recalibration. Their use among academic and community-level NICUs to generate quality reports for benchmarking is widespread.

Three methods dominate, partly as a result of collaboration by NICUs. The first is VON-RA, developed by the Vermont Oxford Network (VON), a collaborative effort of more than 800 NICUs predominantly in the United States. VON developed VON-RA from its database of patient records, of which approximately half are from patients who had a birth weight of less than 1,500 grams. The other two dominant methods are the Score for Neonatal Acute Physiology (SNAP-II) and its related measure, the Score for Neonatal Acute Physiology Perinatal Extension (SNAPPE-II), and the Clinical

Risk Index for Babies (CRIB II). The primary difference among the methods is the risk factors they consider; VON-RA includes some nonphysiological measures (e.g., gestational age, race, presence of a congenital anomaly), while SNAP-II and CRIB-II use primarily physiological parameters (e.g., body temperature, urine output). SNAPPE-II uses both physiological and nonphysiological data, such as birth weight. Critics of SNAP-II and CRIB-II note that the physiological measures used in these methods may be affected by quality of care—when quality of care is what these methods are being used to determine.

Today, VON-RA is widely used to compare NICU performance in real-world clinical settings. Members of VON receive quarterly quality reports that compare their NICU's performance to that of their peers on several measures, including mortality. In addition, they can request comparative data specific to a geographic region or type of institution. VON data have also been used to compare policies and procedures among NICUs and employed in collaborative quality improvement efforts (Horbar 1995; Horbar et al. 2003; Horbar, Soll, and Edwards 2010).

With advances in NICU care, the risk-scoring algorithms must be frequently refined and recalibrated to reflect differences in mortality over time. As part of recent efforts to recalibrate SNAP-II/SNAPPE-II, researchers compared SNAPPE-II and VON-RA and found that they performed similarly when predicting mortality (Zupancic et al. 2007).

PICU Outcomes

Methods for predicting mortality in PICUs have also undergone extensive study. Several risk adjustment methods for PICU patients were developed in the 1990s. The most commonly used tools include the third iteration of the Pediatric Risk of Mortality score (PRISM III) (Pollack, Ruttimann, and Getson 1988; Pollack, Patel, and Ruttimann 1996), the revised version of the Paediatric Index of Mortality (PIM2) (Pearson, Stickley, and Shann 2001), and the Pediatric Logistic Organ Dysfunction (PELOD) score (Lacroix and Cotting 2005; Bestati et al. 2010; Leteurtre et al. 2010). PRISM/PRISM III and PIM/PIM2 have undergone more extensive testing. Both incorporate physiological data and certain diagnoses. PELOD measures four parameters of organ function. PRISM III scores incorporate more data elements than do PIM2 scores, including laboratory data. PRISM III relies on data generated within 24 hours of admission to the PICU, whereas PIM2 uses data generated from the first hour in the PICU. In the same way that quality of care in the NICU could affect the measures SNAPPE-II uses, quality of care in the PICU could affect the physiological and time-dependent measures used by PRISM III. Some argue that this potential problem limits PRISM III's usefulness for comparing quality among PICUs. PRISM III also presents the challenge of how to predict mortality for patients who die within the first 24 hours of a PICU stay (Brady et al. 2006); some of the data elements may be abnormal because the patient is dying, but these data cannot be used to predict death because the patient is already dying.

These risk adjustment methods are incorporated into a continuously updated database of 300,000 PICU patient records, called Virtual PICU Systems. These data have been widely used for risk-adjusted benchmarking and in quality improvement efforts. Researchers also have used PRISM III to study differences in outcomes among patients admitted from referring hospitals and in-house transfers (Gregory et al. 2008), examine differences in outcomes related to pre-PICU resuscitation for shock (Han et al. 2003), and examine supine versus prone positioning of children with acute lung injury (Curley et al. 2005).

Pediatric Hospital Admissions for Congenital Heart Disease and Other Conditions

Surgery to correct congenital heart defects in children is relatively common at regional centers and, like adult cardiac procedures, is technically demanding. In-hospital mortality following pediatric heart surgery varies widely among children and adults, although high-volume institutions generally achieve better outcomes (Jenkins et al. 1995). For these reasons, the outcomes of congenital heart surgery are logical targets for comparing hospital performance (Erickson et al. 2000). Two approaches are taken to adjusting mortality outcomes for congenital heart surgery. Proponents of the first approach argue that risk of death is closely linked to the type of procedure attempted; therefore, the method relies on procedure codes from billing data (Jenkins et al. 1995, 2000, 2002; Hannan et al. 1998). The logic to this approach stems from the observation that the choice of procedure reflects the underlying anomaly, and the cardiac anomaly drives risk of mortality. While this rationale is accurate to some extent, it is also true that congenital heart anomalies in a given phenotypic category are not of equal severity, and cardiac anomalies sometimes coexist with other abnormalities (e.g., kidney or lung problems) that also might contribute to mortality. Furthermore, any risk adjustment method that relies on specific procedure codes needs to be updated frequently to keep pace with advances in surgical technique. Therefore, other investigators have advocated a more traditional approach that first uses clinical physiological data (and then procedure codes or something else entirely). Researchers are also developing child-specific risk adjustment methods for the National Surgical Quality Improvement Program described in Chapter 8 (Morrato, Dillon, and Ziegler 2008).

Common reasons for hospitalization, such as asthma exacerbations, may be sensitive to the quality of outpatient care (Perrin et al. 1989; Billings et al. 1993). In addition, significant numbers of pediatric hospitalizations involve supportive as opposed to curative care and therefore may vary widely in resource use. Because these two factors influence pediatric hospitalization rates and resource use, comparisons of these measures among populations or institutions require risk adjustment.

Among child hospitalizations, asthma admissions and outcomes have received the greatest scrutiny. Risk adjustment of asthma outcomes presents

significant challenges. The most commonly used performance measures—emergency department utilization and anti-inflammatory medication use—are indicators of both severity and quality of care. Some researchers have used prior hospitalization in adjusting for risk, reasoning that the higher threshold for admission to a hospital is more closely related to severity of illness than to quality of care (Ferris et al. 2001). While hospitalizations for asthma are significantly related to asthma outcomes, they do not occur frequently enough among the total population of children with asthma to make them useful as a risk adjuster. Similarly, risk adjusters derived from administrative data, such as ACGs, may have little value in adjusting asthma performance measures (Shields et al. 2002). On the other hand, clinical variables, such as oxygen saturation and respiratory rate, improve predictions of hospital admissions and subsequent short-term outcomes (Ferris et al. 2001; Finkelstein et al. 1995; Homer et al. 1996). These adjustment variables are limited because they control only for severity of acute exacerbation of asthma (appropriate for short-term resource use and outcomes), not the severity of the underlying disease. However, Lieu and colleagues (2002) used parental reports of underlying (chronic) asthma severity as an effective adjustment in comparisons of both longer-term outcomes and quality of care.

Gaps in Pediatric Risk Adjustment Methods

Except mortality adjusters for NICU and PICU care, risk adjustment methods are much less developed for pediatric populations than for adults. However, NICUs and PICUs provide acute, high-technology care, which children rarely need. The majority of care for pediatric populations occurs outside such tertiary settings. As noted earlier, for predicting costs, pediatric components have typically been added to methods covering adults, and statistical performance for pediatric subpopulations generally differs from that for adults. Nevertheless, the distinctly nonrandom distribution of children across clinicians, institutions, and health plans, some of which see disproportionate numbers of socioeconomically disadvantaged children, drives the need for better pediatric risk adjustment. In addition, the experiences of children with little or no health care, and those at risk but not yet diagnosed with specific conditions, need to be better understood.

Given current data limitations, capturing pertinent risk factors for children poses special challenges. More information about the role of families, communities, and local health care environments is needed to understand better the risks these factors pose to the health of children (Halfon and Hochstein 2002; Larson et al. 2008). Crucial insight could be gained by linking information on parents with that of their children (Kahn et al. 2002; Minkovitz et al. 2002; Duncan, Brooks-Gunn, and Klebanov 1994; DiFranza

and Lew 1996). Linking the health data systems (where they exist) of public health departments, private practitioner offices, outpatient clinics, neighborhood health centers, and children's hospitals would require tremendous effort and expense. Instead, several states are merging data from their pediatric public health systems, including, in some cases, Medicaid data. Sharing these data is important to obtaining population-based views of child health, various risk factors, and the unique and overlapping roles of different sources of pediatric care. These observations are not specific to risk adjustment but highlight significant barriers to obtaining comprehensive information about child health, let alone the quality of pediatric care. As is true for adults, risk adjustment for children is inherently limited by inadequate or unavailable data.

A variety of technical issues need additional study. Improved methods for handling small sample sizes are essential for comparing risk-adjusted performance across providers who see children (exclusively and nonexclusively). Changes in patterns of care (e.g., diagnostic testing, therapeutic interventions) and practice differences across settings could affect the data available for risk adjustment and the relationships between risk factors and outcomes of interest. Empirically derived risk adjusters require periodic updating. Finally, like risk adjustment of adults' outcomes, risk adjustment of pediatric outcomes will never be perfect; inclusion of all pertinent risk factors is impossible.

Understanding the effects of excluding potentially important risk factors on comparisons of outcomes across plans or providers is important. Battles over measurement methods should not subvert the resources, attention, motivation, and willpower essential to achieving the ultimate goal of improving health care for children, as has happened in some adult settings.

Notes

1. ICD-9-CM codes are available for some developmental disabilities. Examples include "mild mental retardation" (317, IQ 50–70); "moderate mental retardation" (318.0, IQ 35–49); "severe mental retardation" (318.1, IQ 20–34); "profound mental retardation" (318.2, IQ < 20); and "unspecified mental retardation" (319). Mental retardation is one of few conditions for which ICD-9-CM provides specific clinical definitions (i.e., the indicated IQ levels). Named after an eight-year-old girl, "Rosa's Law," enacted in 2010, required that the term "mental retardation" be changed to "intellectual disability" in all references to the disability in federal law.
2. One possible solution would be to use data on socioeconomic status of populations by zip code or US Census tract as proxies for unavailable sociodemographic information (Fiscella and Franks 2001).

CHAPTER

14

RISK ADJUSTMENT FOR MENTAL HEALTH CARE

Anthony P. Weiss and Mark A. Blais

More than $57 billion is spent every year to provide mental health care to Americans, making it the third most costly—and the fastest growing—segment of the US health care system (Soni 2009). Yet access to mental health care remains inadequate, and the quality of that care is uneven at best (IOM 2006). This paradox, a pointed example of what Kissick (1994) called the "Iron Triangle of Health Care," has generated a series of questions among key stakeholders. Patients want more detail on the quality of clinicians or hospitals. Payers wonder if they are receiving optimal value on behalf of their customers. And mental health professionals contemplate how best to restructure payment arrangements and care delivery approaches to meet the needs of the growing number of people seeking treatment.

Each group of stakeholders seeks data on a key but missing variable in the equation: an accurate measure of the clinical outcomes of mental health care services. Measuring mental health outcomes is complex. Equally complicated is adjusting these outcomes for the intrinsic differences of patients. Risk adjustment methods for mental health outcomes have historically lagged those for the medical mainstream, but the great need to produce meaningful mental health quality metrics will likely drive new developments in this field.

Current State of Mental Health Care Risk Adjustment

Little has changed in risk adjustment for mental health outcomes since Richard C. Hermann wrote the mental health chapter for this book's third edition. The use of large administrative databases (see Chapter 5) as a source of outcomes and risk factor information (Ellwood 1988) has not been realized despite the significant resources expended to compile them. Efforts using data from government payers, such as the Centers for Medicare & Medicaid Services and the Department of Veterans Affairs, have explained only small percentages of the costs and clinical outcomes of care (Hermann, Rollins, and Chan 2007), although some recent adaptations of their models show promise for predicting costs and outcomes of mental health care accurately (Sloan et al. 2006; Rosen et al. 2010). Most approaches are local, limited in scope, and highly specific

to the developer's goals. A good example is development of a risk adjustment model for length of stay on a particular inpatient psychiatric unit (Blais et al. 2003; Hopko et al. 2001). While such efforts may potentially benefit local quality improvement activities, these models are unlikely to generalize to other settings or be of widespread use.

Many unique factors associated with the nature and delivery of mental health care have impeded development of adequate risk adjustment models. Examination of the four core questions raised by risk adjustment (i.e., What outcome? What time frame? What population? What purpose?) in the mental health context highlights these challenges.

What Outcomes Are We Trying to Predict?

Consistent with the broader construct of risk adjustment in health care, most risk adjustment methods for mental health care focus on predicting clinical outcomes or resource utilization (Hendryx, Beigel, and Doucette 2001).

Clinical Outcomes of Mental Health Care

In other medical disciplines, morbidity (often defined physiologically) and mortality are the mainstays of outcomes measurement. However, neither morbidity nor mortality is a key outcome measure for the majority of mental health conditions. As a result, mental health outcomes measurement has relied primarily on assessment of various aspects of mental illness, such as distress, symptom severity, and functional impairments. Typically, mental health outcomes information is based on clinicians' observations or patients' self-reports. Exhibit 14.1 presents a selected list of risk factors drawn from the literature that may be associated with mental health outcomes.

The field's understanding of the neurophysiology associated with mental health–related conditions (including substance abuse) has expanded extraordinarily since the mid-1990s. Nevertheless, the field still lacks biologically based measures that objectively capture the effects of treatment on these neural systems. At some future point, neuroimaging techniques or genetic data may be used for psychiatric disease staging and outcomes measurement (Roffman et al. 2006). Despite their research applications, these technologies are not in routine clinical use.

Mental health conditions can carry a risk of death. Suicide is a mortal outcome for several major mental illnesses and an ever-present consideration in treatment environments. Individual variables are statistically correlated with risk of suicide, including diagnosis of a mood disorder, prior suicide attempts, and increasing feelings of hopelessness. Nonetheless, prospective studies of composite models of suicide risks have not demonstrated sufficient predictive utility to be of clinical or actuarial value (Galfalvy, Oquendo, and

EXHIBIT 14.1 Selected Risk Factors for Predicting Mental Health Outcomes*

Demographic characteristics

- Age
- Race and ethnicity

Clinical factors

- Mental health diagnoses
- Extent and severity of mental health diagnoses
- Number of past psychiatric hospitalizations
- Psychometric test results
- Psychotic symptoms (current and past)
- Mania or hypomania (current and past)
- Current level of dangerousness to self, others, or property
- Need for restraint or seclusion
- Past suicide attempts
- Non-suicidal self-injury
- Past acts of aggression and current aggressive behavior (physical and verbal)
- Comorbid medical conditions and their complexity
- Number of current medications (all types)
- Cognitive functioning (memory, judgment, insight, and ability to communicate)
- Physical pain
- Bladder control
- History of falls

Psychosocial and socioeconomic factors

- Educational attainment
- Marital or partner status
- Degree of social support
- Domestic violence and victim of abuse (emotional, physical, sexual)
- Housing
- Neighborhood characteristics

Health-related behaviors and activities

- Excessive alcohol use (current and past)
- Illicit drug use (current and past)
- Adherence to treatment
- Impulsivity
- Sleep quality
- Functional status (basic and instrumental activities of daily living), including ability to manage medications

Attitudes and perceptions

- Expectation of treatment success

*This list includes selected risk factors that published literature suggests may be related to mental health outcomes. This list is not exhaustive.

Mann 2008). Indeed, most methods have not performed any better than those used in Pokorny's (1983) landmark study of 4,800 veterans over a five-year period in which the positive predictive value of the best available model was less than 3 percent. This performance may reflect the statistical limitations inherent in predicting extremely rare events (Peduzzi et al. 1995); the base rate for completed suicides in the United States is 11.5 per 100,000 person-years (Xu et al. 2010). Expansion to predicting suicidal *behavior* (rather than completed suicide), objectively defined using standardized instruments (Posner et al. 2007), offers promising opportunities to increase predictive accuracy (Hendin et al. 2010).

Without "hard endpoints" of death and physiological morbidity, clinical outcomes for mental health conditions have traditionally been assessed on the basis of symptom severity and overall life functioning. Quality of life has emerged as an important outcome, especially for individuals with severe and persistent mental illness. Mental health outcomes are typically obtained by having patients (via self-report scales) or clinicians (via clinician rating scales) rate pre- and posttreatment symptoms, psychosocial functioning, and other dimensions of interest. Sometimes this rating process is integrated into clinical care (for examples, see Baer and Blais 2009). Hundreds of scales are available for evaluating the outcomes of mental health treatment, but little consensus exists about which measures should be used for which patients. Ongoing debates regarding this important component of treatment reflect in part the complexities of mental health care and also reveal conceptual divisions within this multidisciplinary field. Disagreement around how to define and measure mental health outcomes complicates efforts to develop dynamic health services research programs in this field.

Howard's phase model of change has emerged as one way to guide measurement of mental health care outcomes (Howard et al. 1993, 1996). According to this model, mental health outcomes occur in three separate but related domains: well-being, symptom severity, and psychosocial functioning. Change within the domains follows a predictable pattern: Patients first experience improved well-being (remoralization); symptom reduction follows (remediation); and gains in interpersonal and social functioning occur yet later in treatment (rehabilitation) (Howard et al. 1996).

Many well-validated self-report scales are available to measure well-being and symptom severity, including the Behavior and Symptom Identification Scale-24 (BASIS-24; Eisen et al. 2006), the Brief Symptom Inventory (BSI; Derogatis 1993), the Outcomes Measure-45 (OQ-45; Lambert et al. 1996), and the Schwartz Outcomes Scale (SOS-10; Blais et al. 1999). Multiple disease-specific scales also exist, such as the Patient Health Questionnaire-9 (PHQ9; Kroenke, Spitzer, and Williams 2001) for major depression. These scales are psychometrically sound and well represented in outcomes research. To obtain multiple perspectives on outcomes, researchers have patients

complete self-reports of well-being and symptom severity and have clinicians rate global severity/impairment and psychosocial functioning. The most common measure of global functioning is the Global Assessment of Functioning (GAF; APA 1994). The GAF is a 100-point scale used by clinicians to produce a single rating covering a patient's psychiatric symptoms and social and occupational functioning. The GAF is brief and widely used, and with training, clinicians can produce GAF ratings of adequate reliability (Hilsenroth et al. 2000). However, because the GAF combines multiple dimensions into a single score, its meaning is somewhat ambiguous. Another measure of functioning is the World Health Organization Disability Assessment Schedule (WHODAS-II). The 36-item WHODAS-II measures six domains of functioning and is available in different forms (patient or clinician informant), lengths (12 to 36 items), and languages. The SF-36 Health Survey and the shorter SF-12 also measure physical and social functioning (Ware, Kosinski, and Keller 1994). The SF scales have extensive general and population (disease) specific norms that facilitate benchmarking and cross-study comparisons.

In addition to these measures of actual outcomes, research has shown that patients' expectations and early treatment alliance (defined as the experience of collaboration and connection between patient and provider) are strongly related to various mental health outcomes, including outcomes of psychotherapy (Horvath 2001) and pharmacotherapy (Krupnick et al. 1996). Similarly, patients' pretreatment expectations regarding the effectiveness of care are also powerful predictors of mental health outcomes (Krell et al. 2004; Meyer et al. 2002). These variables are not currently part of routine clinical assessments. Nonetheless, they may have considerable potential as risk factors because they are easy and inexpensive to measure early in treatment.

Despite decades of using measurement in the field of psychology and the prominence of measurement in randomized controlled trials in psychiatry, the use of outcomes measurement for routine mental health practice remains in its infancy (Weiss, Guidi, and Fava 2009). As a result, documentation of psychometric measures in clinical records is limited, and major administrative databases do not contain such measures. This dilemma is critical. As long as mental health treatment relies on proxy outcomes such as psychometric test scores, the best predictor of posttreatment outcomes will almost always be the patient's pretreatment score. It is highly unlikely that any other variable will have a greater correlation with the outcome score than the autocorrelation of the pre- and posttreatment test scores. Furthermore, once the variance associated with the pretest score is removed from the posttest score, the (residual) variance caused by other predictors is usually minimal.

Resource Utilization and Cost Outcomes

In contrast to the complexities of measuring clinical outcomes, measuring resource utilization (most often length of inpatient hospitalizations in days)

and costs (in dollars) is more straightforward. These data are readily available in large administrative data sets, although critical clinical variables may be absent. Data availability may explain why the number of published models examining mental health care utilization and cost outnumbers those measuring clinical outcomes by a five-to-one margin (Hermann, Rollins, and Chan 2007).

Nevertheless, models for predicting mental health care resource utilization and costs are limited: The ability of existing models to prospectively predict mental health–associated costs is essentially no better than existing models' ability to predict clinical outcomes. These inadequacies were partially responsible for the initial exemption of psychiatric hospitals and units from Medicare's original DRG-based prospective payment system (Lave 2003). Decades of subsequent research evaluating alternative case-mix adjustment approaches based on diagnostic codes and readily available demographic data (including Ambulatory Care Groups, Adjusted Diagnostic Groups, and Hierarchical Condition Categories) have produced little improvement. None of the models accounted for more than 10 percent of the variance in expenditures; they generally underestimated costs associated with the presence of a mental health diagnosis and overestimated the costs of care for patients without a mental health diagnosis (Ettner et al. 1998, 2001).

This issue became more significant when the 1999 Balanced Budget Refinement Act mandated implementation of a per diem prospective payment system for psychiatric hospitals and units. The Inpatient Psychiatric Facilities Prospective Payment System (IPF PPS), implemented as a result of this legislation, is the current method for determining reimbursement for inpatient psychiatric care of Medicare beneficiaries. While the IPF PPS does adjust for comorbidities associated with especially high-cost patients, the overall model retains most of the limitations of the DRG-based approaches described in the previous paragraph (Drozd et al. 2008).

At least three factors appear to compromise models' ability to accurately predict costs associated with mental health care. First, most case-mix methods use diagnosis as the primary classification variable, yet as described later in the chapter, the use of psychiatric diagnoses is inherently limited. Second, patients with the same diagnosis may have differing degrees of illness severity, social support, and access to outpatient care, all of which contribute to greater inpatient resource utilization and cost (Drozd et al. 2006). Most standard risk adjustment models do not account for these factors. Third, treatment of psychiatric patients varies widely across providers, resulting in variation in costs incurred. For example, one study found that rates of medication use in the treatment of mood disorders such as major depression and dysthymia varied fourfold across similar clinics (Kramer et al. 2000). Inpatient use of electroconvulsive therapy for treatment of severe major depression also appears to vary significantly (Latey and Fahy 1988) and is associated

with substantially greater resource utilization and cost (Cromwell et al. 2005). These factors may help explain why facility effects (which are not included in standard risk adjustment approaches) may account for up to 50 percent of inpatient psychiatric costs and service utilization (Gifford and Foster 2008).

Over What Time Frame?

As noted in Chapter 4, the time window of observation substantially affects the ability of statistical risk adjustment models to predict both clinical and cost outcomes for most health conditions. But once again, the nature and treatment of mental health conditions make standard approaches to defining this time frame problematic.

Defining the episode of care for a given mental health condition is exceedingly complicated, especially because of evolving views that psychiatric disorders are primarily chronic conditions rather than discrete illness events. For example, many patients with unipolar or bipolar mood disorders have residual symptoms between their diagnosable episodes of mania and depression. These symptoms not only impair persons' functioning but also predict relapses to more severe states (Judd et al. 2008). Even the longstanding concept of the discrete "psychotic break" as the event heralding a schizophrenia diagnosis is being reexamined, as evidence suggests that prodromal aspects of schizophrenia long predate the psychotic symptoms. Furthermore, inter-episode cognitive impairment may be as functionally compromising to these patients as are the delusions and hallucinations most commonly seen in the inpatient setting (Harvey 2010). Defining where to draw temporal lines for outcome measurement related to any of these major mental health conditions is therefore more complex than previously realized.

In addition, the time window chosen for determining outcomes will likely interact with the clinical outcome tool selected. As suggested by Howard's phase model of change described earlier, psychosocial functioning is expected to improve late in the course of treatment. Thus, models assessing the risk of poor functional outcomes will likely yield inaccurate results if a truncated time window is chosen.

Finally, different stakeholder groups may be predisposed toward using different approaches to define episodes of care. For example, clinicians may view an episode of care as the time the patient was undergoing treatment with them; payer organizations may define care in terms of business cycles (fiscal years); and patients may use their return to a previous (premorbid) state of mental health to mark the end of an illness episode. Each definition or perspective has implications for the model and measurement method used for risk adjustment.

In What Population?

Identification of the risk factors associated with the population of interest appears to be a substantial challenge in mental health care. For this reason, current predictive models poorly explain variations in outcomes, even when a well-validated clinical outcome measure and an adequate time frame are chosen. Several considerations relatively unique to mental health care, discussed in the following paragraphs, appear to contribute to these difficulties.

Basic demographic factors that predict risk for many medical conditions (e.g., age) explain little of the clinical or cost-associated risks in mental health care. Key demographic variables have a complex relationship with mental health outcomes. For example, intense psychotic episodes requiring hospitalization may decrease among older patients with schizophrenia.

Most risk adjustment models lean heavily on diagnosis. While diagnoses have some predictive ability, diagnosis-based models in mental health care have more limited explanatory power than those for medical or surgical disorders, likely due to a variety of factors. First, the diagnoses mental health professionals assign to their patients vary significantly, although diagnosis has become more consistent since publication of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Health Disorders* (abbreviated DSM), now in its fourth edition (DSM-IV; at the time of this writing, a fifth edition is under development). DSM-IV specifies the criteria required to formally make a diagnosis, thereby promoting some degree of consistency between the diagnosis and the patient's clinical presentation. Most DSM-IV diagnoses are associated with diagnostic codes from the *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM; see Chapter 5) and then further characterized by indicators of subtypes and clinical severity. For example, the ICD-9-CM code for a depressive disorder (296) can be modified to indicate a recurrent unipolar depression (296.3) or a bipolar depression (296.5). This code can then be further modified to indicate the severity (e.g., mild unipolar depression is 296.31). Unfortunately, most studies exploring the reliability of clinician-generated diagnoses have concluded that as diagnoses become more specific (e.g., beyond three-digit ICD-9-CM codes), interrater reliability becomes unacceptably low (Sartorius et al. 1993).

Second, diagnoses coded on billing forms and actual diagnoses may differ significantly. For example, clinicians may hesitate to include codes for personality disorders or substance abuse conditions on insurance claims (or other administrative forms) when this clinically relevant but potentially socially detrimental information has no bearing on reimbursement or service allocation. Third, clinicians may not adequately document all comorbid conditions (especially comorbid personality disorders [DSM Axis II] and comorbid medical conditions [DSM Axis III]) that have an important effect on the likelihood

that a certain outcome may occur (Morey and Ochoa 1989; Clayton 1990). Fourth, severity of illness is often not clearly documented despite its obvious influence on outcomes. Fifth, important aspects of the psychosocial milieu in which treatment is taking place are often poorly documented and rarely objectively measured. For example, in the DSM-IV, "Psychosocial and Environmental Problems" are addressed by a simple checklist on Axis IV of the multiaxial system (APA 1994). Sixth, treatment refractoriness—the degree to which prior treatments have not benefited the patient—is an important prognostic indicator but is rarely recorded or coded in a consistent manner.

The end result is that available administrative data often fail to represent obvious clinical differences that drive real-world outcomes. The following two cases exemplify this problem:

- Ms. A is 33 years old with a mild recurrence of unipolar major depression (ICD-9-CM code 296, DSM-IV classification 296.31) in the context of a recent downturn in her law practice. One prior episode of depression, while a college freshman, was successfully treated with a low dose of a serotonin reuptake inhibitor and a brief course (six sessions) of cognitive behavioral therapy. She has no comorbid psychiatric concerns and no history of problem substance use. She is physically well. Ms. A lives with a supportive husband and has support from her parents, who live two towns away.
- Ms. B is 33 years old and having her ninth episode of major depression (ICD-9-CM code 296, DSM-IV classification 296.31) following the breakup of another in a long series of tumultuous relationships. She has a history of poor inter-episode symptomatic recovery and chronically poor functional status. Prior to this episode, Ms. B has attempted suicide three times by drug overdose, has four prior psychiatric hospitalizations, and has unsuccessfully tried eight different antidepressants in combination with five other psychotropic medications. She often does not comply with medication regimens due to a combination of excessive alcohol use and limited financial resources. She is morbidly obese and has a history of hypothyroidism, which has been poorly controlled. Ms. B is currently staying at a shelter for battered women and receives limited income from Supplemental Security Income disability payments.

The clinical discrepancies between these two cases are dramatic, but these two individuals, who have widely divergent risk profiles, have been assigned the same diagnosis code. Neither case is unusual in routine clinical practice. To analysts using only available coded data, these cases look identical.

Last, in addition to being less reliable and less clinically descriptive, psychiatric diagnoses are less prescriptive of treatment than are other medical

diagnoses. Therefore, even if clinicians did accurately (reliably) diagnose the type and severity of mental disorders, differences in the nature and intensity of the treatments applied (e.g., weekly psychotherapy, combined medication and psychotherapy) would obscure diagnoses' relationships to outcome variables of interest.

For What Purpose?

As noted at the outset of this chapter, three constituencies—payers, providers, and patients—have compelling interests in better understanding mental health outcomes. However, each group needs risk adjustment for its specific purpose, such as predicting costs, comparing clinical outcomes across providers, or measuring access to patient-centered care. These different goals entail different risk adjustment models or the application of different weights to a common set of variables representing risk factors.

If the goals of using risk-adjusted data lack transparency or are misaligned, data providers may not authorize the use of the data. A cautionary tale is the 2008 attempt by Blue Cross Blue Shield of Massachusetts (BCBSMA) to use a 70-item standardized outcome measurement tool to implement pay-for-performance reimbursement for outpatient mental health care (Liptzin 2009). BCBSMA's lack of clarity around the ultimate intent of the measurement raised concerns that the data might be used for utilization review or provider tiering, despite the initiative's initial billing as a quality improvement exercise. Met with strong provider criticism over the perceived lack of scientific quality, inordinate length, and limited perceived utility of the instrument, BCBSMA abandoned this effort.

Ranking clinicians without performing adequate risk adjustment and publicly reporting the results will stir the ire of mental health professionals, as has happened in other clinical disciplines (e.g., reporting of mortality rates for individual cardiothoracic surgeons). While mental health has lagged other specialty areas in such efforts, initiatives are afoot that may produce the first public reports of mental health care outcomes. In 2009 the National Quality Forum (NQF) certified The Joint Commission's Hospital-Based Inpatient Psychiatric Services (HBIPS) measure. Publishing the results gathered through this measure will be the first major attempt to publicly report mental health care quality, albeit in the specific context of inpatient care. None of the seven measures that compose the HBIPS set will be risk adjusted. Thus, public reports on participating hospitals (which at this time is required of freestanding psychiatric hospitals) will provide unadjusted data on such measures as restraint usage and the number of patients discharged on more than one antipsychotic medication. Similarly, risk adjustment is not proposed for a separate set of mental health measures recommended by the NQF in its National

Voluntary Consensus Standards for Mental Health Outcomes report. In this case, the unadjusted percentage of patients with major depression who have a PHQ9 score of less than five (remission) at 6 and 12 months appears to be a likely candidate for certification. This measure may be problematic for a number of reasons, the most simple of which is that patients with high scores on this scale (indicating more severe depression) at baseline will be less likely to have a score of less than five at follow-up when compared with patients who first present with lower scores (milder depression severity).

Where Do We Go from Here?

Adequate risk adjustment for mental health care outcomes remains elusive. Yet the need for good models to predict mental health care costs, correct clinical outcome inequities in patient populations, and ensure access and patient satisfaction is intensifying. As demonstrated by the implementation of the IPF PPS, the BCBSMA outcomes measurement effort, and the public reporting of unadjusted HBIPS data, stakeholders are moving forward, with or without adequate risk adjustment models. In many ways it is refreshing to see mental health treated in the same manner as other sectors of health care. However, the unintended consequences of these efforts could be disastrous to the mental health care delivery system and its patients. Inadequately compensated capitation plans and public reporting/physician profiling based on unadjusted outcomes data could easily lead to "cherry picking" (i.e., prompt providers to select the "easiest" patients, e.g., Ms. A) and "lemon dropping" (i.e., prompt providers to avoid the most difficult patients, e.g., Ms. B). Significant curtailment of mental health care services could also result and further tighten already perilous access impediments. The prospect of these consequences calls for an immediate solution.

What can be done? In coming years, it will be critical to convene the three major mental health stakeholders to better align their common interests. The clear objective uniting the three is to provide the best possible care to the greatest number of people at a sustainable cost. Better documentation of standard diagnostic information and the routine use (and documentation) of a few validated mental health outcome tools as a part of clinical practice are goals to strive for in the era of electronic health records (EHRs; see Chapter 6). Ensuring the privacy of these data while maximizing their availability and use presents a challenge but is essential. Clinically meaningful information about mental health will need to be shared in a protected and thoughtful way among providers, payers, and patient advocates. Routine measurement, documentation, and rigorous development of risk adjustment models for mental health care will be costly but critical to producing meaningful, actionable assessments of mental health care outcomes. The move to universal

EHRs offers unprecedented opportunities to achieve many of these desired improvements to mental health information quality. However, as is the case for all new technologies, methods and procedures will need to evolve to optimize these capacities. The transfer of current clinical documentation practices to EHRs is unlikely to substantially improve the quality of data or care (Tsai and Bond 2008).

Mental health researchers must continue to seek risk factors that are systematically associated with mental health outcomes across diagnoses, treatment settings, and patient populations. The list of mental health risk factors supported by the literature in Exhibit 14.1 will likely grow as the field gains insight into the genetic and biological bases of mental health conditions and other scientific aspects of these diseases.

CHAPTER

15

RISK ADJUSTMENT AND PERSONS WITH DISABILITIES

Lisa I. Iezzoni

When President George H. W. Bush signed the Americans with Disabilities Act (ADA, P.L. 101-336) on July 26, 1990, people with disabilities assumed the right, in Bush's words, to "pass through once-closed doors into a bright new era of equality, independence and freedom" (Young 1997, 231). Today, roughly 54 million Americans live with disabilities. Despite the ADA and other governmental mandates, however, they are still often left behind, even by the health care delivery system. *Healthy People 2020* (2012), which sets national health priorities for this decade, notes that

> compared with individuals without disabilities, persons with disabilities are more likely to:
>
> - Experience difficulties or delays in getting health care they need.
> - Not have had a mammogram in the past 2 years.
> - Not have had a Pap test within the past 3 years.
> - Not have had an annual dental visit.
> - Not engage in fitness activities.
> - Use tobacco.
> - Be overweight or obese.
> - Have high blood pressure.
> - Experience symptoms of psychological distress.
> - Receive less social and emotional support.
> - Have lower employment rates.

Providing health care services to persons with disabilities also generates enormous expenses. Estimates of disability-associated health care costs for the noninstitutionalized, US civilian adult population in 2006 totaled $397.8 billion (26.7 percent of national health care expenditures) (Anderson et al. 2010, 44). Of the national total, disability-related payments were $118.9 billion for Medicare beneficiaries (38.1 percent of Medicare expenditures), $161.1 billion for Medicaid recipients (68.7 percent of Medicaid costs), and $117.8 billion for privately insured and uninsured individuals (12.5 percent of nonpublic health care expenditures). Persons with certain types of disabilities (e.g., conditions requiring considerable medical intervention or community

support services) are likely to generate higher costs from year to year than are persons with other types of disabilities, making capitated health plans or programs such as accountable care organizations (ACOs) less motivated to enroll them in their panels. Incentives for organizations like ACOs to enroll high-risk, high-cost individuals are central to current proposals for controlling escalating Medicare costs (Rosenthal 2011). "CMS should consider an approach to risk adjustment [for capitation payments] that would create incentives to keep and recruit those very patients and allow the ACO to concentrate on patients with the greatest potential for savings, which would benefit both the program and the ACO" (MedPAC 2011b, 8). These high-risk populations likely include many persons with disabilities.

Furthermore, in many situations, the most effective care for persons with disabilities is difficult to determine. With relatively few exceptions, randomized controlled trials of medical treatments explicitly exclude persons with significant disabilities. Therefore, studies of treatment effectiveness for persons with disabilities rely on observational studies (see Chapter 11). Analysis of observational data generally requires risk adjustment, as noted throughout this book, but risk-adjusting outcomes for persons with disabilities is complex, starting with fundamental questions about the definition of disability. In addition, outcome studies of persons with disabilities must reach beyond standard medical concepts, such as the primacy of diagnosis and characteristics inextricably bound to individuals, and consider the critical role of societal and environmental factors.

This chapter introduces major conceptual and practical issues raised by considering disability in risk adjustment. Section 4421 of the 1997 Balanced Budget Act (P.L. 105-33) mandated Medicare prospective payment of inpatient rehabilitation facilities (IRFs). Starting in January 2002, CMS required IRFs to report clinical information using the IRF-Patient Assessment Instrument to risk-adjust Medicare's prospective IRF payments. This risk classification method built on the Functional Independence Measure (FIM) and function-related groups (FRG), which have been extensively described in the literature (Stineman and Granger 1997; Stineman et al. 1994, 1997; Berlowitz and Stineman 2010). This chapter addresses these measures briefly. Chapter 16 addresses topics relating to long-term care, including nursing homes, which often care for individuals with disabilities.

Defining Disability

Defining disability is complex because of its multilayered personal, institutional, administrative, programmatic, and societal ramifications. Since the fourteenth century, disability has delineated categories of people meriting societal assistance: alms, food, and shelter. However, "because physical and mental incapacity are conditions that can be feigned for secondary gain . . .

the concept of disability has always been based on a perceived need to detect deception" (Stone 1984, 23). Starting in the nineteenth century, physicians were charged with differentiating deserving from undeserving disabled persons, making these distinctions using theoretically objective findings from clinical examinations and diagnostic technologies. The traditional medical model of disability assumes that individuals "afflicted" with compromising health conditions must adapt their lives and expectations to their limitations.

Over the last half century, various definitions of disability have been developed for diverse purposes. Some definitions echo the traditional medical model, whereas others introduce a new concept: Disability results from social and physical environments that fail to accommodate persons with differing physical, sensory, cognitive, or emotional abilities. People are not disabled; society is (Shapiro 1994; Oliver 1996; Charlton 1998). This social model spurred demands for accommodations that would enable people with disabilities to participate fully in daily life throughout communities and workplaces. No single definition, however, serves all societal purposes. Depending on the context, definitions of disability differ widely, as suggested by the examples taken from the ADA, the Social Security Administration (SSA), and the World Health Organization's (WHO) *International Classification of Functioning, Disability and Health* (ICF) in Exhibit 15.1.

EXHIBIT 15.1 Examples of Disability Definitions

Source	Purpose	Disability Definition
Americans with Disabilities Act	Civil rights protections	Section 3: "(A) a physical or mental impairment that substantially limits one or more of the major life activities . . . ; (B) a record of such impairment; or (C) being regarded as having such an impairment"
Social Security Administration	Determining eligibility for income support	"Inability to engage in any substantial gainful activity by reason of any medically determinable physical or mental impairment(s) which can be expected to result in death or which has lasted or can be expected to last for a continuous period not less than 12 months"
World Health Organization	Classification system	"Umbrella term for impairments, activity limitations or participation restrictions," conceiving "a person's functioning and disability . . . as a dynamic interaction between health conditions (diseases, disorders, injuries, traumas, etc.) and contextual factors," including environmental and personal attributes

Sources: 42 USC § 12102(2); SSA (2008); WHO (2001).

The causes, nature, timing, pace, and societal implications of disabling conditions are diverse. Some conditions are congenital; others are acquired. Some occur suddenly by injury or accident, while progressive chronic conditions arise slowly. Some gradually limit but do not threaten life; others hasten death. Some are visible to outsiders; others are hidden. Some engender stigmatization and blame; others prompt pity and paternalism. Some are primarily considered diseases (e.g., cancer, heart failure, emphysema) but can become profoundly disabling.

Perceptions of disability also differ across individuals and cultural groups. For example, many people who are born deaf speak American Sign Language and do not view themselves as disabled; they participate fully in a distinct Deaf culture. (This cultural identification will possibly change with the growing use of cochlear implants in children who are born deaf.) According to the 1994–1995 National Health Interview Survey Disability supplement, almost 20 percent of manual wheelchair users do not see themselves as disabled, although mainstream society probably does (Iezzoni et al. 2000b). Others see a compromised life, while persons with disabilities do not. Morris (1996, 62) interviewed a woman named Ruth Moore, whose spine was "crumbling," risking complete paralysis. Moore worried about her physicians' attitudes, observing:

> The neurosurgeon told me that he was only interested in quality of life and that in no way would he be looking to prolong my life if he didn't feel the quality would be acceptable. However, neither he nor anyone else has asked me what criteria *I* would use in judging what was an acceptable quality of life. I am very worried that if I get admitted unconscious or without the power of speech, he will take a decision based on *his* judgment and *his* criteria about what is an acceptable quality of life.

With civil rights and public debates about disability access and accommodations came semantic sensitivities. Language matters; *crippled*, *lame*, and *gimp* are clearly out (although some people with disabilities use this language to make certain points), as are *handicapped* and *challenged*. The commonly used phrases "confined to a wheelchair" and "wheelchair bound" convey an image of someone lashed into place, an inaccurate perception (wheelchair users do get out of their chairs). Some, for example, drive cars or ski. The author Nancy Mairs, who uses a power wheelchair, does not view herself as immobile. "Relaxed and focused, I feel emotionally far more 'up' than I generally did when I stood on two sound legs. . . . Certainly I am not mobility impaired; in fact, in my Quickie P100 with two twelve-volt batteries, I can shop till you drop at any mall you designate, I promise" (Mairs 1996, 38, 39). Wheelchair users today advocate active rather than metaphorical language: They *use* a wheelchair; they are not confined to or by their wheelchair.

This discussion has three important implications for studies and policies involving persons with disabilities. First, the population of interest must

be clearly defined in precise, descriptive language. However, the data source used for analysis may constrain potential definitions of populations with disabilities (see discussion later in the chapter). Second, findings generalize only to populations defined using comparable criteria. Finally, the answer to the "Risk of what?" question (see Chapter 2) may vary as widely as the definitions of disability shown in Exhibit 15.1 (e.g., level of physical functional impairments, ability to work, quality of life). Standard biomedical and psychosocial frameworks for outcome measurement may be insufficient for specifying outcomes meaningful to persons with disabilities (Stineman and Qu 2007). A broader, "biopsycho-ecological" perspective would better recognize the (Stineman and Streim 2010, 1035)

> complex interacting multilevel functional hierarchies beginning at the cellular level and ending at the individual's experience of the environment. Although the foci of illness and injury are within the body and mind, the physical and social environments contain elements that can cause or exacerbate disease and barriers that interact in ways that lead to injuries and disabilities.

Risk Factors for Assessing Disability Outcomes

As just noted, for risk adjustment of outcomes among persons with disabilities, the answer to the question "Risk of what?" may be complicated and multidimensional. For some purposes, such as setting capitation payments for a year of care, the answer is straightforward. The Medicare Program for All-Inclusive Care of the Elderly (PACE) enrolls beneficiaries aged 55 or older who are sufficiently disabled or medically needy to be deemed "nursing home certifiable." As described in Chapter 7, to set capitation payments for PACE, the algorithm considers a frailty adjuster that captures enrollees' functional status (Kautter, Ingber, and Pope 2008; Hirth, Baskins, and Dever-Bumba 2009).

As outlined in Chapter 8, the first step in developing a risk adjustment approach involves specifying conceptual models linking potential risk factors to pertinent outcomes. To do so, developers must consider the chosen definition of disability. For example, in the SSA's employment-centric disability definition, all pertinent risk factors relate to the individual's characteristics; it ignores the possibility of workplace accommodations. In contrast, the ICF's definition encompasses a full range of risk factors (see Chapter 3), recognizing that some risk factors arise from individuals, whereas others arise from the person's physical and social environments.

Drawing from theoretical frameworks such as the ICF, rehabilitation outcomes may range from the functioning of specific organs, structures, and body systems to the participation of persons in homes, social networks, and communities. Pertinent risk factors vary across outcomes. For instance,

research investigating the range of motion around a particular joint following an intervention would consider fewer risk factors than would a study examining posttreatment ability to perform daily activities.

Risk-adjusting rehabilitation outcomes is more difficult than risk-adjusting outcomes of other services, such as stays in intensive care units (ICUs). In ICUs, clinicians typically control therapeutic interventions (e.g., intravenous medications, oxygen flow); these treatments are easily quantified and carefully calibrated to specific physiological indicators, and the outcome is often obvious (e.g., mortality). In contrast, rehabilitation interventions are difficult to quantify, especially outside institutional settings. Furthermore, outcomes depend on myriad factors, including not only patients' physical and cognitive abilities but also their underlying medical diseases, willingness to participate actively in their care, and the existence of supportive physical and social environments. As described later in this chapter, patients' preferences for different outcomes are especially critical. One study asked persons with various medical conditions and functional limitations to imagine recovery from 15 different types of deficits; imagined difficulties with toileting and with depression troubled participants the most (Stineman et al. 2007). Participants made trade-offs among domains of physical, psychological, and social functioning; some valued physical independence over psychological well-being or social abilities, while others displayed opposite patterns. These preferences can affect patients' engagement in rehabilitation and adherence to regimens that may take weeks, months, or longer to complete or may require critical changes to their lifestyle or home environment.

Role of Diagnoses

People with disabling conditions are not necessarily sick. For example, persons who have been deaf or blind since early childhood may be in superb physical and mental health, requiring only routine, preventive care. Even if a specific etiology is found, the cause of their deafness or blindness may carry few implications for their future health. Rehabilitation services, for instance, generally involve patients who are physiologically stable; vital signs and internal organ function are controlled and are not at immediate risk of serious decompensation. Measures of physiological stability are therefore less relevant in most rehabilitation contexts, with the possible exceptions of pulmonary or cardiac rehabilitation. Nevertheless, other standard medical measures, such as diagnoses, are potentially important risk factors, depending on the population of interest and other contextual factors.

In many instances, a disease, disorder, or condition underlies the impairment or functional limitation. In children and younger adults, disabilities usually arise from single conditions, while in late middle-aged and older adults, multiple coexisting conditions (multimorbidities) frequently produce the functional deficits. Multimorbidities may also be related to other factors that might

affect disability (or the ability of individuals to accommodate disability), such as overall lifetime earnings (Tucker-Seeley et al. 2011). In these circumstances, identifying a single underlying diagnosis may be neither possible nor useful. Secondary conditions caused by the impairment or functional problem, such as pressure ulcers, urinary tract infections, and incontinence, may also become risk factors.

In some contexts, coexisting conditions are less important than the nature of the impairment or functional limitation. For example, in deriving version 2.0 of the FIM-FRGs, Stineman and colleagues (1997) found that including multiple diagnoses did not significantly improve the measure's ability to predict length of rehabilitation hospital stays, although their findings were limited by the inadequacies of their coded diagnostic data. Another study found that for persons with the most severe functional deficits, the number of comorbidities was not significantly associated with the risk of developing medical complications during inpatient rehabilitation (Siegler, Stineman, and Maislin 1994). In contrast, for persons with milder functional impairments, comorbidities significantly predicted complications. Berlowitz and colleagues (2008) examined the ability of three commonly used comorbidity measures—the Charlson Comorbidity Index, Adjusted Clinical Groups (ACGs), and Diagnostic Cost Groups (DCGs)—to predict outcomes of stroke rehabilitation six months after admission to VA facilities. The models had different predictive performance; of the three, DCGs produced the highest c-statistics.

Cognitive Ability, Mental Health, and Sensory Functioning

Cognitive ability and mental health (e.g., depression, anxiety, fear) are potentially important risk factors for disability outcomes, for various reasons. For example, when rehabilitation services require active understanding and participation by patients, cognitive functioning and mental health may influence treatment intensity and patients' adherence to the regimen. In addition, these attributes may reflect debility in general and the likelihood of further progression of functional impairments. A review of 78 studies examining functional status declines among elderly persons living in the community found significant associations with cognitive impairment and depression (Stuck et al. 1999). Intact cognitive status is especially predictive of recovery of activities of daily living (ADLs) among older persons (Gill, Robison, and Tinetti 1997; Hardy and Gill 2004).

Depression and anxiety are significantly associated with functional dependence and secondary conditions, such as falls and incontinence (Tinetti et al. 1995), although the causal pathways may be circular: Depression and anxiety may produce functional dependence, which exacerbates depression and anxiety, and so on. Patients who are depressed or discouraged generally participate less actively in their care (Kane 1997). Although these mental health problems are important risk factors, they can often be treated effectively;

however, clinicians frequently neglect or avoid confronting these conditions among patients with disabling conditions. Some clinicians may believe that depression is inevitable among these patients, an opinion consistent with broader societal stigmatization of debility and mental health problems. Especially in studies of outcomes for quality assessment, untreated mental health problems could represent not only risk factors but also substandard care.

Cognitive and mental health factors are also important predictors of resource consumption in certain contexts. Cognitive FIM scores significantly predicted length of stay for 5 of the 21 FIM-FRG impairment categories (stroke, traumatic brain, lower extremity fracture, joint replacement, and major multiple trauma with brain/spine injury) (Stineman et al. 1997). Alzheimer's disease patients (largely defined by impaired cognition) generate substantially lower costs to Medicare for the next year of care than do persons without this disease (Ellis et al. 1996; Ash et al. 2000), perhaps because Alzheimer's patients receive fewer expensive, intensive services. Until 2002, Medicare generally refused to pay for rehabilitative services for Alzheimer's patients, arguing that rehabilitation provided little benefit.

Sensory function (or dysfunction) could represent comorbid conditions (e.g., macular degeneration, age-related hearing loss). However, separate consideration of these often neglected conditions highlights their importance. For example, among persons aged 65 or older, 23 percent of those without mobility difficulties reported having had vision tests, compared with 22 percent of those with major mobility problems (Iezzoni et al. 2000a). In this age group, however, 26 percent with major mobility problems had serious difficulty seeing, even when using glasses or contact lenses, compared to only 5 percent of those without mobility problems. Poor vision and hearing problems are major risk factors for falls and further functional declines (Tinetti et al. 1995; Cassel, Besdine, and Siegel 1999). A review of 78 studies found that visual deficits were strongly associated with other functional declines among elderly persons living in communities (Stuck et al. 1999). Sensory difficulties may also contribute to mental health–related risk factors. For example, hearing loss is associated with isolation, confusion, and depression among elderly patients (Lachs et al. 1990).

Self-neglect, especially among older patients, is increasingly recognized as either a sign of or a risk factor for functional declines (Dyer et al. 2007; Pavlou and Lachs 2008). This syndrome crosses various domains, including physical, cognitive, social, and economic. Persons who exhibit self-neglect are also particularly susceptible to experiencing abuse, which can exacerbate disability.

Sociocultural Factors, Preferences, and the Physical Environment

Especially once patients leave institutional settings, their social, home, and community environments and other personal factors affect their rehabilitation

outcomes. An extensive literature documents the association between worse health and socioeconomic disadvantage, including poor education, poverty, unemployment and low-income occupations, and social isolation (Secretary's Advisory Committee 2008). These same factors contribute to disability through different pathways.

Health beliefs, including patients' perceptions of control over their health outcomes, strongly influence patients' willingness to adhere to rehabilitation regimens (Merrill 1994; Chen et al. 1999). Patients' preferences and sociocultural attitudes are especially important in contexts involving assistive technologies, such as mobility aids like walkers and wheelchairs (Iezzoni 2003). Patients' and their families' negative views regarding assistive devices influence patients' willingness to use them, even when the potential benefits appear obvious to clinicians. Mobility and seating aids prompt particularly negative views, perhaps because they reflect, "in a tangible and objective fashion, the visible reality of the increased dependencies" (Gitlin, Luborsky, and Schemm 1998, 174). Methodological, logistical, and other factors complicate efforts to study the effects of assistive technologies and thus their potential contribution as indicators of risk (Hoenig, Giacobbi, and Levy 2007). New technologies, such as the use of geolocation monitoring, are assisting in documenting usage patterns of wheelchairs and other mobility aids (Sonenblum et al. 2008).

The physical environments in which patients live and conduct their daily activities may significantly affect disability (Tabbarah, Silverstein, and Seeman 2000; Wahl et al. 2009; Clarke and Nieuwenhuijsen 2009). Whether these environments physically accommodate persons with functional deficits and enhance their safety are key factors. Socioeconomic circumstances, such as the financial wherewithal to renovate homes or move to more accessible surroundings, are also intertwined. Personal preferences and other social forces, such as household composition and interpersonal relationships, influence patients' willingness to alter their home environments to improve accessibility and safety. The presence of certain features in community environments, such as neighborhood parks and walking areas, may also affect disability (Beard et al. 2009; White et al. 2010).

Functional Status, Disability, and Overall Health

Prior functioning is the best predictor of future functioning. Therefore, most studies of outcomes for people with disabilities use some measure of baseline functioning or global performance (e.g., ADLs) as key risk factors. The literature on functional measures and measurement is extensive (see Chapter 3). Some approaches aim to predict specific outcomes (e.g., the FIM-FRGs predict rehabilitation hospital costs), whereas others are more generic. Measures for children may differ from those for adults. Even theoretically "objective" measures of function vary by context. *Capability* indicates what a person "can do" in controlled settings, whereas *performance* assesses what a

person "does do" in everyday life. Capability typically exceeds performance (Young et al. 1996).

Functional status may also be a critical risk factor for other outcomes (see Chapter 3). For example, Palsbo and colleagues (2010) used patients' reports to create a disability case-mix indicator that has four clusters of activity limitations. They then compared enrollees' ratings of three Medicaid plans in California using the Consumer Assessment of Healthcare Providers and Systems survey (CAHPS; see Chapter 7). They found that disability produced stronger biases in relative CAHPS ratings of health plans and specialists than did demographic factors.

This broad class of potential risk factors forces an important question: Whose perspective should drive this assessment, the patient's or the clinician's? The answer depends on the context, especially because patients' and clinicians' assessments of functional status and ultimate functional goals may diverge (Stineman et al. 1998; Rist et al. 2008). Ensuring the accuracy of FIM assessments has been a challenge to maintaining the integrity of IRF prospective payment. Research suggests that a variety of factors might contribute to inaccurate FIM measurements by clinicians (Wolfson, Doctor, and Burns 2000; Doctor et al. 2003). Although many measures of patient impairment and functioning are clinicians' assessments, the patient's voice is now widely recognized as critically important. Depending on the context, the best strategy may be to obtain functional status information directly from patients.

Examination of health-related quality of life, self-perceived activity limitations, and life satisfaction recognizes the validity and importance of the perspectives of persons receiving rehabilitation (Whiteneck 1997; Kramer 1997). These measures distill the myriad social and environmental factors affecting disability into a specific and intrinsically meaningful metric: how persons with disabilities feel. People who perceive their quality of life as poor may be less motivated to participate actively in their care and less likely to do well, for diverse reasons. For many purposes, health-related quality of life may be the target outcome, as well as an important risk factor.

Administrative Data and Identifying Disability

Administrative databases do not link information about health conditions with insight into disability (e.g., performance of daily activities, participation in life situations, social and physical environmental barriers) (Iezzoni 2002). Most administrative data report only diagnoses and procedures (see Chapter 5). Certainly, some diagnoses infer potential physical disabilities, as shown by the ICD-9-CM codes in Exhibit 15.2. Diagnoses alone, however, generally convey little about their effects on people's daily activities or the effect of

EXHIBIT 15.2 Examples of ICD-9-CM Codes Representing Physical Functional Impairments

7993	Debility, unspecified
438	Late effects of cerebrovascular disease
3420	Flaccid hemiplegia
3421	Spastic hemiplegia
3429	Hemiplegia, unspecified
3440	Quadriplegia
3441	Paraplegia
3442	Diplegia of upper limbs
3443	Monoplegia of lower limb
3444	Monoplegia of upper limb
34481	Other specified paralytic syndromes, locked-in state
3449	Paralysis, unspecified
34460	Cauda equina syndrome without mention of neurogenic bladder
34461	Cauda equina syndrome with neurogenic bladder
V440	Tracheostomy
V441	Gastrostomy
V460	Aspirator
V461	Dependence on respirator
V468	Other enabling machines
V469	Unspecified machine dependence
V538	Wheelchair

social or physical environments. Palsbo and collaborators (2008) found that adding Healthcare Common Procedural Coding System claims and number of prescription medications to diagnoses improved the ability of their administrative data–based algorithm to correctly classify individuals reporting functional limitations. Whether ICD-10-CM will improve the identification of persons with disabilities from administrative data is unclear at this time.

Administrative data can identify people meeting SSA definitions of disability: persons eligible for Medicare and Medicaid through Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI), respectively. These important populations have significant health and health care costs and concerns (Anderson et al. 2010). My colleagues and I used Medicare disability eligibility for persons under age 65 to demonstrate disparities in breast and lung cancer care and outcomes between these individuals and persons under age 65 without Medicare (McCarthy et al. 2006; Iezzoni et al. 2008a, 2008b). However, findings from SSDI and SSI recipients may not generalize to other persons with similar disabling conditions, who for whatever reason have neither applied for SSDI/SSI nor qualified as disabled under SSA provisions.

Claims or Encounter Records

To generate claims or encounter records and their associated diagnosis and procedure codes, patients must have a health insurance benefits package that covers the specific item or service. Despite social safety net programs, notably Medicaid, some persons with disabilities are uninsured: From 2000 to 2006, according to our analyses of federal survey data, 11.6 percent of working-age persons with disabilities did not have health insurance, compared to 14.8 percent of all working-age individuals (Iezzoni, Frakt, and Pizer 2011). Persons with disabilities were especially likely to lack insurance (22 percent) in southern states that had particularly stringent Medicaid eligibility requirements (Pizer, Frakt, and Iezzoni 2009). Even among the uninsured, persons with disabilities were at a disadvantage: 36.0 percent of uninsured persons with disabilities reported being unable to obtain necessary medical care, compared to 9.5 percent of uninsured, nondisabled persons, and 26.9 percent of uninsured persons with disabilities reported being unable to obtain necessary medications, compared to 5.3 percent of uninsured individuals without disabilities (Iezzoni, Frakt, and Pizer 2011). Those who are unable to obtain services have no trail of claims.

The health insurance benefits package limits coverage of services that address long-term functional deficits and environmental factors contributing to disability. For instance, Medicare explicitly excludes coverage for "any services that are not reasonable and necessary for the . . . diagnosis or treatment of illness or injury or to improve the functioning of a malformed body member" (42 C.F.R. Sec. 411.15[k]). This medical necessity criterion especially limits reimbursement for assistive devices, such as wheelchairs, eyeglasses, and hearing aids. Furthermore, services obtained from outside sources or reimbursed out of pocket are not included in health insurance databases. Thus, the items and services represented by claims or encounter records do not fully reflect the health needs of people with disabilities.

Diagnosis Codes

Exhibit 15.2 includes examples of ICD-9-CM codes representing physical functional impairments; numerous other codes depict low vision and deafness, mental retardation, cognitive impairments, mental health problems, medical diseases, and other conditions underlying disability. Creative combinations of ICD-9-CM codes produce clinical stories suggestive of a person's disability, functional status, and daily life. For example: Patient XYZ has multiple sclerosis (code 340) and weakness of the legs (code 344.9, paralysis unspecified), which is a "condition influencing [his] health status" (code V49.2, motor problems with limbs). The person uses a wheelchair (code V53.8), has an inaccessible home (V60.1, inadequate housing), and is unemployed (V62.0).

Health services research studies have used diagnosis codes from claims to identify populations with chronic conditions and disabilities at risk for high

future costs of care (Kronick, Zhou, and Dreyfus 1995; Kronick et al. 1996, 2000). By definition, however, diagnosis codes in administrative databases are not intended to tell clinical stories about patients; they are intended to be used to generate reimbursement. In particular, V codes, a potential source of insight about both disabling conditions and environmental barriers (e.g., inadequate housing), are coded infrequently, and insurers rarely base reimbursement on V codes.[1] As noted earlier, how the advent of ICD-10-CM will affect this situation remains to be seen.

Procedure Codes, Durable Medical Equipment, and Outpatient Pharmacy

Some procedures imply debility or potentially disabling physical impairments, such as amputation of a limb, hip arthroplasty, knee replacement, transplantation of a major organ, insertion of a tracheostomy tube (for long-term mechanical ventilation), or placement of a gastrostomy tube (for feeding). Certain treatments, such as chemotherapy and radiation therapy, also suggest potentially disabling conditions. People receiving such interventions often generate high health care costs (Pope et al. 2000b).

Medicare, Medicaid, and private insurers generally pay for acute care services, except in cases when preapproval is required but was not obtained. Therefore, administrative databases from fee-for-service insurance plans supposedly represent almost all procedures or treatments obtained within the period they cover (see Chapter 5). However, reliance on procedure codes to identify people with disabilities raises important considerations. First, people receiving these procedures are a nonrandom subset of all people with disabling conditions. Second, even for potentially life-prolonging interventions (e.g., gastrostomy tube placement), significant variations in procedure use are likely because of differing preferences for care, practice styles, and service availability across patients, physicians, institutions, and geographical regions. Third, persons may have received the service prior to the period covered by the administrative database. Although certain V codes indicate prior procedures (e.g., transplant status), pertinent V codes are not always listed. Fourth, because of restrictive payment policies limiting physical, speech-language, or occupational therapy, only certain patients receive these services, so drawing inferences about disability on the basis of the presence or absence of these interventions is problematic. Fifth, the presence of procedure codes reveals little about outcomes—that is, whether the intervention improved functioning. A knee replacement, for example, may eliminate a patient's mobility difficulties and hence potential disability.

Durable medical equipment (DME) and assistive technologies frequently compensate for or ameliorate functional limitations and can prolong persons' ability to live independently in their homes and communities (Brummel-Smith and Dangiolo 2009; Ganesh et al. 2011). DME and assistive

technologies are diverse, ranging from supplemental oxygen to mechanical ventilation devices, parenteral and enteral nutrition supplies, limb prostheses, wheelchairs, walkers, hospital beds, grab bars, and ramps. New technologies enable remote monitoring of persons for falls and exacerbations of health conditions (Kang et al. 2010). One risk adjustment method uses codes for DME to capture risks related to disability (van Kleef and van Vliet 2010). Challenges to this approach include instances in which insurance does not cover equipment or disputes a patient's need for certain types of equipment and persons have to acquire the devices through other means (e.g., out-of-pocket payments, charitable donations).

Identifying people with disabilities by DME claims thus raises questions. Most important, as noted earlier, restrictive coverage policies mean that paid claims for DME (especially mobility aids) represent only a fraction of patients' need. Medicare and private insurers often reimburse certain DME only in limited circumstances (e.g., for homebound patients, persons who must use the equipment in their homes, or persons who have not obtained similar equipment in the previous five years). In addition, like procedure use, DME use varies significantly on the basis of differing personal preferences and practice styles. People may have obtained DME before the period covered by the database, and V codes (e.g., indicating wheelchair use) are reported sporadically. Although claims for DME offer useful information, they do not identify all people with disabling conditions who use assistive technologies.

As noted in Chapter 5, administrative databases from Medicare Part D, Medicaid, the VA, and private insurers offering prescription drug benefits contain outpatient pharmacy claims data. Pharmacy data help identify people with specific conditions, such as diabetes, bipolar disorder, asthma, and HIV infection. These data suggest the severity of medical conditions: For example, people receiving insulin presumably have more refractory diabetes mellitus than do those using oral hypoglycemic agents. Certain drugs suggest the pace of illness or the course of disease: For instance, multiple sclerosis patients prescribed interferon-beta probably have relapsing-remitting rather than secondary progressive disease. However, pharmacy data raise cautions identical to those prompted by procedure and DME codes. They identify only certain subgroups of people; varying practice patterns compromise the generalizability of findings; and prescriptions not covered by insurance plans escape detection.

Merging administrative data with other data sources can significantly enhance the analytical utility of these files for disability-related research, as discussed in Chapter 5 (Iezzoni 2002). For example, in the longitudinal Medicare Current Beneficiary Survey (MCBS), the answers selected for certain questions create disability indicators. The MCBS oversamples Medicare beneficiaries under age 65 and contains population sampling weights to produce national estimates (www.resdac.org/MCBS/data_available.asp).

International Classification of Functioning, Health and Disability

Although the ICF (WHO 2001) classifies functional abilities and social and environmental contexts, this coding scheme is rarely used in the United States. In contrast, European nations have initiated concerted efforts to use the ICF to generate information on population disability and develop core code sets for specific disabling conditions (Almansa et al. 2011; Cerniauskaite et al. 2011; Kostanjsek 2011). They view the ICF as an indication of an important shift of WHO away from its traditional focus on infection control to an increasing recognition of the need to reduce the burden of health problems worldwide (Cieza and Stucki 2008, 303):

> Avoiding both sociological and biomedical reductionism, the ICF provides a starting point for a comprehensive and integrative understanding of the universal human experiences of functioning and disability, where body, personal action and the overall physical, social and attitudinal environment in which the person lives and acts are inextricably intertwined.

WHO views the ICF as complementing its *International Classification of Diseases;* each classification offers insight into different aspects of health. The ICF groups its alphanumeric codes into 30 chapters under four broad headings: body functions (physiological functions, including cognitive and psychological functions), body structures (anatomical parts of the body, such as organs, limbs, and their components), activities (execution of tasks or actions by individuals) and participation (involvement in life situations), and environmental factors (the physical, social, and attitudinal environment in which people live and conduct daily life). Qualifiers added to ICF codes indicate the extent or magnitude of an impairment in body function or structure (from no impairment to complete impairment) and the level of difficulty a person has performing activities or participating in daily life situations (from no difficulty to complete difficulty). Environmental factors may serve as either a barrier or a facilitator, indicated along a scale ranging from none to complete.

The following vignette demonstrates how ICF codes alongside diagnosis and procedure codes can portray complex clinical scenarios and potentially enhance the information retrievable from administrative databases considerably (Iezzoni and Greenberg 2003): Patient ABC, in his mid-40s, was diagnosed with multiple sclerosis (ICD-9-CM 340) several years ago. He now has increasing difficulty walking (ICD-9-CM 719.7, ICF d4509.3), but he can move around with only mild difficulty (ICF d465.1) using a motorized scooter (ICF e1201.+3). He retired from his job as a school bus driver (ICF d8452) and has recently qualified for SSDI (ICF e5700.+4). Patient ABC's health insurance

has lapsed (ICF e5650.4), and he is not yet eligible for Medicare.[2] He is having trouble paying for his medication (ICF d870.2), which includes drugs for multiple sclerosis, hypertension (ICD-9-CM 401.9), and elevated cholesterol (ICD-9-CM 272.9). Multiple sclerosis has also severely affected his vision, impeding him from reading newspapers or books (ICD-9-CM 369.9; ICF b210.3, d920.4). He lives alone in a two-story house, spends much of his day watching television, and frequently feels sad (ICD-9-CM 311, ICF b152.3). He also feels cut off from his old friends (ICF d7500.4).[3]

Despite the potential to gain insights from comprehensive coding using the ICF, considerable practical challenges and some conceptual questions remain before the nomenclature will be accepted in the United States. In particular, distinctions between activity and participation codes remain blurred, and codes for personal factors are not yet fully developed (Whiteneck 2006; Guralnik and Ferrucci 2009; Jette 2009). Other concerns include (1) the absence of a clear crosswalk between ICF language and the familiar concepts of standard ADLs and instrumental ADLs and (2) the lack of a dynamic model of how different attributes and factors contribute to disability (Freedman 2009; Guralnik and Ferrucci 2009). Gerontologists in particular have been reluctant to accept the ICF for these reasons and because of the lack of extensive experience with/testing of the classification scheme. Acceptance of the ICF elsewhere in the world, however, is strengthening calls for greater usage of the ICF in the United States so that information on disability can be comparable worldwide (Jette 2009). The 2011 *World Report on Disability* by WHO and The World Bank (2011, 45) called for worldwide adoption of ICF:

> Using the ICF, as a universal framework for disability data collection related to policy goals of participation, inclusion, and health will help create better data design and also ensure that different sources of data relate well to each other. The ICF is neither a measurement tool nor a survey instrument—it is a classification that can provide a standard for health and disability statistics and help in the difficult task of harmonizing approaches across sources of disability data.

Surveys

The most detailed information on population disability comes from federal surveys (see Chapter 7, in particular Exhibit 7.1). Chapter 7 describes special concerns about accommodating respondents with disabilities so they have equal opportunities to respond to surveys (Palsbo et al. 2011). Defining the question sets by which disability is identified in federal surveys has generated considerable discussion (Freedman et al. 2004; National Research Council 2009). Variations in trends of population disability, particularly in older age, across federal surveys prompted extensive efforts to determine consistent ways to capture self-reported disability. At this time, the ACA (see Chapter 1) is

leading the agenda regarding disability-related questions in federal surveys. Motivated by the desire to gather and report information on health care disparities, Section 4302 of the ACA contains the following language:

> (1) IN GENERAL.—The Secretary shall ensure that, by not later than 2 years after the date of enactment of this title, any federally conducted or supported health care or public health program, activity or survey (including Current Population Surveys and American Community Surveys conducted by the Bureau of Labor Statistics and the Bureau of the Census) collects and reports, to the extent practicable—
>
> (A) data on race, ethnicity, sex, primary language, and disability status for applicants, recipients, or participants;
>
> (B) data at the smallest geographic level such as State, local, or institutional levels if such data can be aggregated;
>
> (C) sufficient data to generate statistically reliable estimates by racial, ethnic, sex, primary language, and disability status subgroups for applicants, recipients or participants using, if needed, statistical oversamples of these subpopulations; and
>
> (D) any other demographic data as deemed appropriate by the Secretary regarding health disparities.

Spurred by this mandate, in June 2011 the Office of Minority Health (OMH 2011) proposed a set of six questions for collecting information on disability in federal surveys. After reviewing submitted comments, OMH released the final list of six disability questions on October 31, 2011 (see Exhibit 15.3). The questions represent five key functional domains or "basic actions"—seeing, hearing, cognition, mobility, and self-care—and are modeled

EXHIBIT 15.3
Data Standard for Disability Status*

1. Are you deaf or do you have serious difficulty hearing?**
2. Are you blind or do you have serious difficulty seeing, even when wearing glasses?
3. Because of a physical, mental, or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions? (5 years old and older)
4. Do you have serious difficulty walking or climbing stairs? (5 years old and older)
5. Do you have difficulty dressing or bathing? (5 years old and older)
6. Because of a physical, mental, or emotional condition, do you have difficulty doing errands alone such as visiting a doctor's office or shopping? (15 years old and older)

*This set of questions is the standard for collecting federal population survey data on disability. The proposed questions appeared in the *Federal Register* (Vol. 76, No. 126, page 38396) on June 30, 2011, and were finalized following a public comment period on October 31, 2011 (http://minorityhealth.hhs.gov/templates/content.aspx?ID=9228&lvl=2&lvlID).

**Responses to each question are *yes/no*.

on questions tested elsewhere in the world (WHO and The World Bank 2011).[4] As stated on the OMH (2011) website:

> The six-item set of questions used by ACS[5] and other major federal surveys to characterize functional disability is proposed as the minimum standard for collecting population survey data on disability. The question set was developed by a federal interagency committee and reflects how disability is conceptualized consistent with the International Classification of Functioning, Disability, and Health. The question set went through several rounds of cognitive testing and has been adopted in most major federal data collection systems.

Regardless of what happens to the ACA, the need to collect a consistent set of disability information across federal surveys to track disparities in care will persist. This information will also be useful for other studies relating to population disability.

Conclusions

Consideration of issues relating to persons with disabilities will become increasingly important in the coming decades. As the Institute of Medicine noted in *The Future of Disability in America*, the number of persons living with disabilities in the United States will grow substantially in the next 30 years primarily because of aging baby boomers (IOM Committee on Disability in America 2007). At the other end of the life span, the number of children and youth living with disabilities also is increasing. Trends linked to childhood and young-adult obesity suggest that disability numbers might also increase during middle age (Alley and Chang 2007; Manton 2008). Thus, in coming years, a substantial fraction of the US population will have one or more disabling conditions. Development of risk adjustment methods that capture disability in a valid, unbiased manner will be critical for a variety of purposes, including setting capitation payments and monitoring quality of care across providers and health care delivery systems.

Notes

1. Exceptions include V581 (maintenance chemotherapy) and V580 (radiotherapy session), which must accompany hospital claims for these services.
2. Persons who qualify as eligible for SSDI become eligible for Medicare 24 months after first receiving cash benefits (which happens 5 months after qualifying for SSDI). Thus, new SSDI beneficiaries must wait 29 months (5 months + 24 months) to qualify for Medicare. The only exceptions are amyotrophic lateral sclerosis patients, who since 2001

immediately qualify for Medicare upon receiving SSDI. Prior to 2001, many amyotrophic lateral sclerosis patients died or became significantly impaired during the two-year wait.

3. ICF d4509.3 = walking, unspecified, severe difficulty; ICF d465.1 = moving around using equipment, mild difficulty; ICF e1201.+3 = assistive product and technology for personal indoor and outdoor mobility and transportation, substantial facilitator; ICF d8452 = terminating a job; ICF e5700.+4 = social security services, complete facilitator; ICF e5650.4 = economic services, complete barrier; ICF d870.2 = economic self-sufficiency, moderate difficulty; ICF b210.3 = seeing functions, severe impairment; ICF d920.4 = recreation and leisure, complete difficulty; ICF b152.3 = emotional functions, severe impairment; ICF d7500.4 = informal relationships with friends, complete difficulty.
4. The 2011 WHO/World Bank report about disability around the world discusses the development of six questions for capturing basic disability concepts (seeing, hearing, mobility, cognition, self-care, and communication) by a worldwide working group. These questions are similar to the US questions (Exhibit 15.3).
5. The ACS (American Community Survey) is an ongoing survey conducted by the US Census Bureau to gather population information at the community level.

CHAPTER

16

RISK ADJUSTMENT FOR LONG-TERM CARE

Dan Berlowitz and Orna Intrator

Long-term care consists of "an array of health care, personal care, and social services generally provided over a sustained period of time to persons with chronic conditions and with functional limitations" (IOM 2001b, 27). It differs from other types of health care in that the goal of long-term care is not to cure illness but to maintain an optimal level of functioning. For many people, the care spectrum begins with acute care for a new condition, transitions to post–acute care for ongoing treatment and rehabilitation, and ends with long-term care for any remaining impairment or disability (Kramer 2002). While informal caregivers such as family and friends provide substantial amounts of long-term care, public and private institutions and organizations also play essential roles.

The rationale and basic framework for risk adjustment of long-term care outcomes parallel those in other care settings, as described in other chapters of this book. However, risk adjustment for long-term care raises important conceptual and methodological issues that differ from those raised by acute and routine outpatient care. This chapter examines risk adjustment for long-term care, emphasizing how it differs from risk adjustment for other health care settings, and describes current applications of risk adjustment to quality improvement. The examples in this chapter primarily involve nursing homes and home health care, the two settings that have received the greatest attention. Risk adjustment for other long-term care settings and populations is still immature. Although many of the same methodological issues apply, relevant outcomes and predictors of these outcomes vary considerably depending on the specific long-term care setting and population studied.

Impetus for Long-Term Care Risk Adjustment

Long-term care risk adjustment is generating considerable interest because of the convergence of four forces: (1) the increasing heterogeneity and complexity of an aging population and its rising demand for long-term care, (2) the growing variety of available long-term care alternatives, (3) concerns about the costs of long-term care and payment policy implications, and (4) ongoing concerns regarding the quality of care.

An estimated 11 million Americans currently need long-term care (Kaye, Harrington, and LaPlante 2010). This highly diverse population consists predominantly of community residents, many of whom are under age 65 and receive informal help. In addition, 1.4 million frail elderly live in nursing homes on a given day, and over 3.5 million spend some time in nursing homes in the course of a year. Due to the aging of the baby boomers, elderly persons are the fastest growing segment of the US population. By 2050, individuals aged 85 or older, the "oldest old," will number over 19 million, up from 5.7 million in 2010 (US Census Bureau 2010). Assuming no changes in age-specific rates of disability, the demand for long-term care services will soar. The number of people needing long-term care is expected to increase by 30 percent over 15 years beginning in 2004, with even more dramatic increases after 2020 (Friedland 2004). Estimates suggest that the number of nursing-home residents will double, and might even triple, by 2030 (AOA 2011).

Demand for different types of long-term care services may depend not only on age but also on race. While overall nursing home use rates have declined over time, white patients have predominantly driven this change. Use among black individuals increased steadily from 1974 (28 per 1,000 elderly black persons) until around 1999 (56 per 1,000) (NCHS 2010a). Nursing home rates for black and white individuals converged in the mid-1990s, breaking the long-standing historical pattern of lower nursing home use among elderly black individuals (Bishop 1999; NCHS 2010a; Ness, Ahmed, and Aronow 2004; Smith et al. 2008). Feng and colleagues (2011) report that between 1999 and 2008, the number of elderly Hispanics and Asians living in US nursing homes grew by over 54 percent, while the number of elderly black residents increased by nearly 11 percent and the number of white residents declined by about 10 percent. Although these shifts have been driven by changing demographics, the number of minority residents in nursing homes has increased more rapidly than has the number of minorities in the general population. These results may suggest unequal minority access to often preferred home- and community-based long-term care services.

Smaller families and growing rates of childlessness will decrease the pool of informal caregivers in future decades, increasing reliance on formal or paid providers (Wolff and Kasper 2006). Whether the workforce of paid long-term care providers will be sufficient to meet the escalating demand is unclear. A more detailed understanding of the aging population, its expected health care needs, and its care preferences will be critical to managing what will likely be increasingly scarce long-term care resources.

Long-term care settings include nursing homes, community-based residential care, and assisted-living facilities, as well as people's homes, where care is coordinated by home health and hospice care agencies. Within these settings, the types of care provided vary widely. Specialized programs are

developed for people with specific needs. For example, rates of hospice use in nursing homes more than doubled between 1999 and 2006 (Miller et al. 2010; Gozalo et al. 2008). Nursing homes also may have dedicated beds or units for residents with dementia. In association with disease advocacy organizations, a skilled nursing facility specializing in care for persons with amyotrophic lateral sclerosis (Lou Gehrig's disease) or profoundly disabling multiple sclerosis opened outside Boston in 2010. The facility also contains residential units for persons with these conditions. New models of nursing home care, such as the Eden and Green House programs, aim to provide care that is more resident-centered and are growing in popularity (Kane et al. 2007).

A better understanding of the cost and quality implications of these changes relative to other settings is essential (Miller et al. 2010; Gozalo et al. 2008). Consumers, regulators, managers, and clinicians are placing increasing importance on understanding and improving care provided in these myriad settings and arrangements. Many audiences therefore seek information that is risk-adjusted by methods specific to long-term care.

As the demand for long-term care escalates, expenditures will rapidly increase. Home health care expenditures, which totaled $64.7 billion in 2008 (including $26.6 billion from Medicare and $22.4 billion from Medicaid), are projected to more than double to $153.8 billion by 2019. In 2008, nursing home expenditures totaled $138.4 billion (including $25.7 billion from Medicare and $56.3 billion from Medicaid) and are expected to increase to $245.9 billion by 2019 (Truffer et al. 2010).

To promote efficiency and control costs, the Centers for Medicare & Medicaid Services (CMS) has increasingly relied on risk adjustment–based payment strategies for long-term care (see Exhibit 16.1). In response to rapidly increasing Medicare expenditures for skilled nursing facilities, in 1998 CMS implemented case mix–based payment based on Resource Utilization Groups (RUGs) III for Medicare beneficiaries in nursing homes (CMS 2012). Periodic refinements to RUGs III aimed to facilitate more accurate payments to nursing homes. On August 8, 2011, CMS presented the final rule on the adoption of the 66-group RUGs IV (*Federal Register* 2011). Many Medicaid programs also now use prospective or mixed (combined prospective and retrospective) reimbursement systems to pay nursing homes (Feng et al. 2008). Similarly, Medicare's home care reimbursement system based on Home Health Resource Groups (HHRGs), called the Home Health Prospective Payment System (HH PPS), was implemented in 2000; refinements to the HH PPS system were issued in August 2007 and became effective in January 2008 (CMS 2011a). New approaches to reimbursement of long-term care are also being developed as Medicaid funding is increasingly used to provide long-term care in noninstitutional settings. Furthermore, risk adjustment is essential to setting appropriate capitated payment levels for CMS programs such as Special Needs Plans and the Program of All-Inclusive Care for the

Elderly to improve efficiency and integration of care (Konetzka and Werner 2010).

Central to prospective payment systems are risk adjustment methods that allocate resources to providers on the basis of demographic and clinical characteristics of patients. To incentivize providers to care for patients requiring the most resources, CMS makes higher payments to providers caring for sicker patient populations (Feng et al. 2008). However, this payment approach is problematic because CMS could inadvertently reward poor quality by paying more for patients whose functional status declines because of inadequate care (Butler and Schlenker 1989). In fact, the financial pressures on nursing homes resulting from decreased average payments and prospective payment

EXHIBIT 16.1 Medicare's Current Risk Adjustment Methods for Nursing Home and Home Health Care Payments*

Method	Payment Episode	No. of Groups	Description
Resource Utilization Groups IV	Day	66	Assigns patients to one of nine categories: 1. Rehabilitation plus extensive services 2. Rehabilitation 3. Extensive services (e.g., ventilator support, parenteral feeding) 4. Special care high 5. Special care low 6. Clinically complex (e.g., oxygen therapy, chemotherapy, terminal illness) 7. Impaired cognition (impaired short-term memory, decision making, orientation) 8. Behavioral problem (daily inappropriate behavior) 9. Reduced physical functions (high ADL dependence) Additional subdivisions are based on ADL deficits, depression, and receipt of nursing rehabilitation.
Home Health Resource Groups	60 days	80	Scores patients on each of three domains: 1. Clinical severity (four levels) including diagnoses, wounds, elimination patterns, intravenous therapy 2. Functional status (five levels) including six ADLs 3. Service utilization (four levels) including recent hospitalization or rehabilitation

*These approaches change over time. The CMS website has the latest updates.

may be prompting nursing homes to provide worse quality of care so they receive greater reimbursement (Konetzka et al. 2006). While many questions remain about the effect of these reimbursement systems on quality, access, and costs of care, limited data suggest that the adoption of case mix–based payments has increased severely impaired Medicaid residents' access to care (Feng et al. 2006).

Concerns about the quality of long-term care persist despite that over 25 years have passed since the Institute of Medicine's (IOM Committee on Nursing Home Regulation 1986) landmark report *Improving the Quality of Care in Nursing Homes* motivated the major regulatory reforms of the Omnibus Budget Reconciliation Act of 1987 (OBRA). Nursing home practices have changed significantly, placing stronger emphasis on resident-centered care, and studies have demonstrated some improvement in the quality of care (Mor et al. 2011a; IOM 2001b; Berlowitz et al. 2000; Hawes et al. 1997). Nonetheless, care often remains suboptimal. Well-publicized cases of flagrant abuse and governmental investigations have fueled concerns about nursing home quality and government efforts to improve it (GAO 2010). Quality problems have also been identified in home health care, where the lack of oversight of health workers is particularly worrisome (Rosati 2009). Work force concerns, such as problems recruiting and retaining employees who receive low wages, overlay worries about the quality of nursing home and home health care.

Market-based reform, including public reporting of performance data and pay for performance, are viewed as central to improving long-term care quality (Konetzka and Werner 2010). Performance monitoring based on valid, reliable, and timely data is critical to these efforts (IOM 2001b). As in other settings, risk adjustment is an essential component of many quality measures. Studies have demonstrated that clinically and statistically credible risk adjustment methods have been developed for long-term care, that different providers see very different populations of patients, and that risk adjustment alters judgments of provider performance (Berlowitz et al. 1996a; Mukamel 1997; Arling et al. 1997; Mukamel and Brower 1998; Porell and Caro 1998; Rosen et al. 2001b; Mukamel et al. 2008). Risk adjustment is also important in determining whether efforts to improve the quality of long-term care are succeeding, as improved outcomes may be due to improved care or differences in patient mix (Berlowitz et al. 2001a). Thus, risk adjustment is necessary to understanding temporal trends in care.

Despite the many factors promoting risk adjustment methodologies specific to long-term care, an overarching long-term care risk adjustment approach may become less practical in the future. The US health care system currently functions within silos, where each component of care is responsible only for its own practice, leading to conflicting incentives and excess costs. However, the advent of private long-term care insurance and the possibility of future health care reforms place a greater emphasis on the need to examine

the actuarial implications of lifetime need for long-term care. As long as silos of care persist, such as nursing homes and home health care, lifetime long-term care utilization will be extremely difficult to contain. Various complex incentives will continue to conflict among these settings. Risk adjustment under these circumstances is exceedingly difficult methodologically, particularly in terms of understanding and accounting for underlying sources of risk.

Data for Risk Adjustment in Long-Term Care

As in other settings, data for risk adjustment in long-term care come from various sources, including administrative databases, medical records, and patients' reports. However, as noted in Chapter 5, administrative databases in long-term care settings, particularly services reimbursed by Medicare and Medicaid, offer much richer clinical insight than do standard, ICD-9-CM code-based acute care hospital or outpatient data sets (Berlowitz, Brandeis, and Moskowitz 1997). Much of the risk adjustment work in long-term care has relied on these specialized administrative databases, such as the Minimum Data Set (MDS), which was conceived in response to OBRA. The 1997 Balanced Budget Act carried significant new ramifications for data reporting and documentation, prompted by fundamental changes in payment policies, notably implementing prospective payment for home health agencies and skilled nursing facilities, as well as inpatient rehabilitation facilities and long-term care hospitals. In particular, Medicare currently requires collection of the following data:

- MDS, now in its third revision (MDS 3.0), was developed specifically to ensure uniform resident assessment systems in nursing homes. MDS data are required to be collected on all residents of nursing homes receiving Medicare or Medicaid reimbursement. MDS includes more than 300 items, including cognitive and sensory function, mood and behavior, physical functioning, performance of daily activities, and other health factors (Morris et al. 1990; Morris, Murphy, and Nonemaker 1995). Version 3.0 was designed to improve the reliability and accuracy of collected data, incorporate patient-reported preferences and satisfaction, increase the accuracy of assessments, and facilitate linkages with electronic health records. MDS is the basis for calculating nursing home prospective payment using RUGs (Fries et al. 1994; Swan and Newcomer 2000) and quality indicators (Zimmerman et al. 1995) and supports CMS's annual certification reviews of nursing homes. CMS requires that MDS data be routinely transmitted to its repository.
- The Outcome and Assessment Information Set (OASIS) for home health care visits includes more than 100 items, including living arrangements,

supportive assistance, sensory and physical functioning, emotional and behavioral status, equipment management, and other factors (Deitz et al. 2010; Shaughnessy et al. 2002). OASIS data are the basis for calculating home care prospective payments using HHRGs. OASIS data also support risk-adjusted quality measurement and reporting in Outcome-Based Quality Improvement (OBQI) reports and surveillance of agency-level potentially avoidable event tracking in the Outcome-Based Quality Monitoring (OBQM) reports.

MDS, OASIS, and other long-term care databases are structured differently from hospital databases because they must account for patients receiving care over extended periods. Administrative databases for acute care hospitalizations capture all information about a hospitalization (e.g., admission and discharge dates, discharge diagnoses, procedures, discharge disposition) in the associated record. In contrast, long-term care databases are typically cross-sectional, describing patients' status on prespecified dates that are not usually related to specific "events." MDS data are collected at fixed times, including at admission, quarterly (every three months), and at times of significant changes in health status. Since data are available only for these prespecified, discrete points in time, most do not reflect the actual dates of any clinical changes.

The contents of long-term care databases also differ from the clinical contents of data sets containing ICD-9-CM diagnosis codes. As described earlier, long-term care databases generally provide more comprehensive information about patients' clinical situations, including detailed descriptions of their ability to perform activities of daily living (ADLs). Databases compiled from MDS and OASIS may contain dozens of clinical indicators; however, information on diagnoses and procedures may be limited. MDS emphasizes solely the reporting of "active" diagnoses—that is, those that affect residents' functioning, treatments, or prognoses. Reliance on assessments of patients' functional status by care providers raises questions about the accuracy and reliability of the data produced by nursing homes, as noted by government reports (MedPAC 2001; GAO 2002). Some argue that the reliability of the data is well established (Mor et al. 2011b; Berg et al. 2002; Morris et al. 1990, 1997; Hawes et al. 1995). More information on data accuracy is necessary to assess the credibility of these data sets, especially with the recent mandate for the extensively revised MDS 3.0.

Databases used for risk adjustment in long-term care have been developed for three primary purposes: regulatory monitoring, reimbursement, and quality. An example of a prominent database developed for regulatory monitoring in nursing homes that is now also widely used in research on quality is the national Online Survey, Certification, and Reporting System (OSCAR). OSCAR includes limited clinical information, and most of the information it does include is facility-level rather than on individual residents. Studies using

OSCAR to examine the association of provider characteristics and risk-adjusted outcomes must therefore rely on facility-level variables describing patient characteristics and outcomes. Examples of such studies include investigations that used the percentage of ADL-dependent residents as a risk adjuster for prevalence rates of pressure ulcers (Zinn, Aaronson, and Rosko 1993), research that used quality inspection surveys to determine institutional deficiency rates (Harrington et al. 2001), and investigations that used the percentage of residents with dementia or psychiatric diagnoses as a risk adjuster for rates of antipsychotic medication use (Hughes, Lapane, and Mor 2000). The use of facility-level data, such as those available from OSCAR, for risk adjustment may sometimes bias conclusions, a situation analogous to the ecological fallacy in epidemiological research. For example, in one study, rankings of nursing homes based on risk-adjusted pressure ulcer rates differed depending on whether patient-level data were used or these same data were first combined into facility-level measures (Fries, Morris, and Skarupski 1998). Additionally, the quality of OSCAR data raises many concerns (Ray 2000; Feng et al. 2005). Thus, while OSCAR remains the main source of data on institutional characteristics, facility-level data on patient outcomes and risk factors from OSCAR will probably be increasingly replaced by resident-level MDS data.

The second reason for developing long-term care databases is to support case mix–based reimbursements. Examples of databases developed for this purpose include those created by the Department of Veterans Affairs (VA) and New York State for RUGs II (Schneider et al. 1988). These databases contain data elements necessary for the RUGs II grouper and therefore include limited clinical information. With the advent of comprehensive databases such as MDS and OASIS, there is limited need for databases that solely provide information for reimbursements.

The final purpose for long-term care databases is to support clinical management, quality assessment, and improvement efforts. For example, Resident Assessment Protocols, which use MDS items to identify the presence of conditions such as delirium or change in cognitive status, were developed to help clinicians identify major issues arising in the care of nursing home residents and to offer guidelines (based on some of the collected data) on how to address them. MDS version 3.0 reverses the process of assessment and care planning. It uses Care Area Triggers, which identify areas of potential issues and conditions that need additional assessment and review and then suggest resources the facility can use to assess whether the issue/condition merits further attention (CMS 2012). Thus, these types of databases usually contain the detailed clinical information necessary for managing and evaluating patients' outcomes and are sufficiently detailed to support development of complex risk adjustment models (Berlowitz et al. 2001b; Berg et al. 2002).

In contrast, for home care, OASIS was not designed to improve the assessment process. Instead, from the outset, OASIS was intended to support a reporting system that could inform outcome-based quality improvement (Shaughnessy, Crisler, and Schlenker 1998; Shaughnessy et al. 2002). Initial design of OASIS anticipated development of risk adjustment methods to assist outcome comparisons.

Although each record is cross-sectional, the strength of long-term care databases is their ability to detect changes in health status over time by linking multiple records of individual patients. Thus, within a given period, an index assessment describes the baseline status of specific patients, which can be linked to later (outcome) assessments. Those outcome assessments then may serve as the index assessment for a subsequent period (see Exhibit 16.2). When building risk adjustment models, developers may use changes in health status between the index and outcome assessments as outcome events, for which characteristics from the index assessment serve as predictor variables (i.e., risk factors). Examples of outcome events include new occurrence of a pressure ulcer in patients without previous ulcers, changes in ADL function, and development of contractures (Arling et al. 1997; Berlowitz et al. 1996a; Mukamel 1997; Shaughnessy et al. 2002). However, the exact date when the change occurred is not known.

Calculation of the rates of these change events inevitably does not factor in some cases. For example, pressure ulcers are not counted if they develop and then heal or if the patient is discharged before the next cross-sectional assessment. Thus, the calculated rates are not "true" incidence rates, although this term is commonly used to describe these numbers.

Each database presented earlier is useful for examining risk-adjusted outcomes in nursing homes or home health care settings. Linkage of several of these databases offers important additional insight (Lipowski and Bigelow 1996). Numerous studies have combined OSCAR and MDS data, enabling

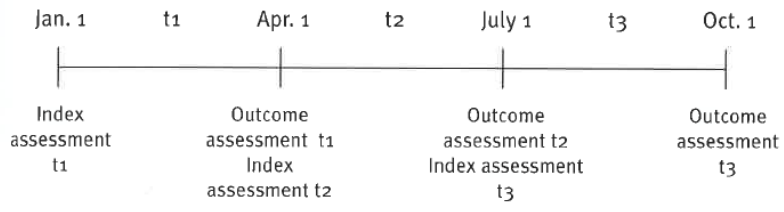


EXHIBIT 16.2 Effects of Cross-Sectional Minimum Data Set (MDS) Assessments

This exhibit shows how cross-sectional MDS assessments performed at 90-day intervals can be used to observe changes in health status in a patient over time. This patient could contribute three observations to the sample data set used to develop a risk adjustment model.

researchers to examine facility characteristics associated with risk-adjusted outcomes. The Residential History File was recently created to also include Medicare Denominator and claims files (Intrator et al. 2011). Medication data are becoming increasingly available from a variety of sources, including Medicare Part D, and will provide additional details on comorbidities and disease severity for developers to incorporate in risk adjustment models. The Systematic Assessment of Geriatric Drug Use via Epidemiology database is an early example that not only contained MDS and OSCAR data but also combined these data with information on medications and Medicare utilization (Bernabei et al. 1999).

Outcome Measures

As noted throughout this book, a fundamental first step in defining risk is specifying the outcome—that is, risk of what? Because of prospective payment, dollars are perhaps the most commonly used outcome for nursing home and home health care risk adjustment. Both RUGs and HHRGs use payment levels to define risk. However, as elsewhere in health care, using clinical outcomes to judge quality is important for long-term care services (Kane 1998).

No global outcome measure captures all aspects of long-term care. Nevertheless, the detailed clinical and functional status information contained in patient-level long-term care databases enables construction of various outcome measures that can provide a broader view of quality than is generally possible with other administrative data sources, such as acute care hospital discharge abstract files. An expansive view that includes both process outcomes and health outcomes is helpful. *Process outcomes* or measures reflect specific care items and utilization associated with patients' well-being and may include hospitalizations, emergency room treatment, or physical restraint use. *Health outcomes* include functional status change, continence, pressure ulcers, dehydration, contractures, and falls. Good performance on one of these outcome measures does not necessarily guarantee good performance on other measures, suggesting the need to examine multiple outcomes (Mukamel 1997; Berlowitz et al. 2005). A composite measure that combines several outcome measures into a single score might be desirable, in terms of efficiency. Nonetheless, it remains unclear whether a composite measure, if created, would represent a single underlying nursing home quality construct (reflective) or simply a combination of uncorrelated items (formative) (Shwartz, Burgess, and Berlowitz 2010). Issues regarding the weighting of these different components also must be resolved.

Although the use of multiple outcome measures is desirable, the specific outcomes must be tailored to the unique situation of the population and

setting being studied. Outcomes relevant to the frail elderly may not be appropriate for children receiving long-term care for chronic disabilities. Special considerations arise when measuring the quality of end-of-life care, to which patients' preferences and expectations are central (Mukamel et al. 2012; Donaldson and Field 1998). Satisfaction and quality of life are increasingly emphasized by resident-centered approaches to care and may be useful outcome measures for many long-term care patients. Studies examining the quality of nursing home care have found important differences among nursing homes in satisfaction and quality-of-life measures (Schnelle et al. 1999; Kane et al. 2004), but these measures could be difficult to capture for populations with high degrees of cognitive impairment. Mortality, which is widely used to evaluate acute hospital care, may be less relevant to long-term care settings and in particular to hospice care because death often represents the final stages of disease progression, not the quality of care. Ideally, studies using mortality as an outcome also incorporate other clinical outcomes and explicitly consider patients' preferences.

Selection of relevant outcome measures depends on the long-term care setting. Functional decline is often studied in nursing homes, while home care typically considers improvement. The literature reaches little consensus on these choices. Even within a setting, relevant outcome measures may be defined differently; one study identified six different definitions of ADL decline in nursing home residents that have been used in research projects (Rosen et al. 1999). The percentage of patients exhibiting ADL decline and judgments of facility performance varied considerably, depending on which of these definitions was used. Measures may also need to consider whether incidence or prevalence rates are used as the outcomes. For example, both incidence and prevalence rates of pressure ulcers have been used as outcome measures. Prevalence rates are generally easier to calculate and comprehend; however, prevalence rates may include cases in which the condition developed outside long-term care settings, so they do not represent current quality of care. This concern does not apply to incidence rates.

Functional status is usually described by ADLs or instrumental ADLs (IADLs). Each activity measure, such as dependence in toileting, feeding, or walking, may be examined individually, or they may be combined into an overall index. Changes in ADL or IADL function may be viewed as continuous variables, or some dichotomous threshold may be set (either people worsen or they do not). If ADL/IADL function is dichotomized, the focus may be on whether or not patients decline, stabilize, or improve. To create scales from multiple ADL and IADL questions, researchers often simply count the number of activity limitations (e.g., zero, one, or more than one ADL difficulty). Finch, Kane, and Philip (1995) developed a summary numeric ADL and IADL scale by asking experts to assign weights to different activities and levels of limitations (Kane et al. 1996, 2000). The scale including only

ADLs ranges from 0 to 5,350, while the scale combining ADLs and IADLs extends from 0 to 6,614 (0 indicates optimal functioning). Because continuous scores weight ADLs and IADLs and do not make dichotomous judgments about impairments, they provide more information than do counts of ADL or IADL difficulties.

Given the many considerations inherent in specifying an outcome measure, selecting outcome measures linked to best practices seems to be the best strategy. Studies have demonstrated that many existing nursing home outcomes–based quality indicators, including those for pressure ulcers, urinary incontinence, and weight loss, do not differentiate among nursing homes performing well and those performing poorly on key processes of care (Schnelle et al. 2003; Bates-Jensen et al. 2003; Simmons et al. 2003). Another study found that a quality improvement initiative conducted in over 30 nursing homes led to a significant decline in the incidence of stage 3 and 4 pressure ulcers but not to improvements in other pressure ulcer measures, suggesting that incidence of pressure ulcers is a more sensitive indicator of quality than is prevalence of pressure ulcers (Lynn et al. 2007).

Risk Factors for Long-Term Care Outcomes

Risk adjustment models for long-term care must account for important patient characteristics that increase the probability of the outcome of interest and that are likely to be distributed unevenly among providers. Some of this uneven distribution may arise from selection (e.g., when a provider "cherry picks" its patients).[1] These characteristics encompass the range of risk factors described in Chapter 3. The relative importance of these different dimensions of risk for predicting outcomes differs between long-term and acute care settings, even for identical outcomes (e.g., costs of care). In addition, significant predictors vary across long-term care outcomes. Therefore, development of appropriate risk adjustment models across the many outcomes and settings of long-term care is a daunting task.

In long-term care, functional status is typically the most important risk factor. Many different diseases, such as cardiorespiratory conditions, cancer, dementia, and musculoskeletal conditions, cause functional deficits. In long-term care, the specific underlying disease is often less important than the extent of functional impairment for predicting outcomes. Consequently, risk adjustment methods for both clinical and cost outcomes of long-term care generally rely heavily on functional status measures.

The limited role of diagnostic data in long-term care risk adjustment is well documented. ADLs are important components of RUGs and HHRGs, while RUGs ignore diagnoses entirely, and HHRGs consider only diabetes, neurological disorders, and orthopedic conditions. For clinical outcomes,

initial risk adjustment models developed for nursing homes generally did not include diagnoses, largely because of data limitations (e.g., the minimal clinical content of the reimbursement databases used in these efforts) (Berlowitz et al. 1996a; Mukamel 1997).

As databases including diagnostic information have become increasingly available, more comprehensive evaluations are now possible. Usually, as anticipated by clinical judgment, only a few diagnoses are significantly associated with specific outcomes. For example, diabetes, peripheral vascular disease, and recent hip fracture are the only significant diagnostic predictors of pressure ulcer development (Berlowitz et al. 2001b). Broader measures of comorbidity, such as number of diagnoses, are not associated with pressure ulcer development. Schizophrenia is associated with declining mental status but not with incontinence, possibly because these patients tend to be younger (Porell et al. 1998). Rosen and colleagues (2000) found that adding specific diagnoses, such as neurological diseases, minimally improved predictions of functional decline, with c-statistics increasing from 0.66 to 0.68.

While measures of functional status are important in predicting outcomes, there is little consensus on how to represent ADLs as **predictors** in a risk adjustment model for a clinical outcome. Many of the considerations described earlier regarding functional status as an outcome are also relevant when it is a predictor. A first decision is often which ADLs to include. For example, the MDS includes ten ADLs, along with measures for bathing and continence. These variables are highly correlated so that paradoxical results may arise if models include too many ADLs. Generally, selection of specific ADLs for inclusion in risk adjustment models should rely on plausible clinical linkages with the outcome (e.g., immobility causes pressure ulcer development). Close examination of the statistical performance of models that include different combinations of ADLs is also helpful. Alternatively, researchers may use an ADL scale, as described earlier (Finch, Kane, and Philip 1995; Kane et al. 1996, 2000), or one developed specifically for the MDS (Morris, Fries, and Morris 1999). Researchers could then compare the statistical performance of risk adjustment models including individual ADLs to that of models using ADL scales.

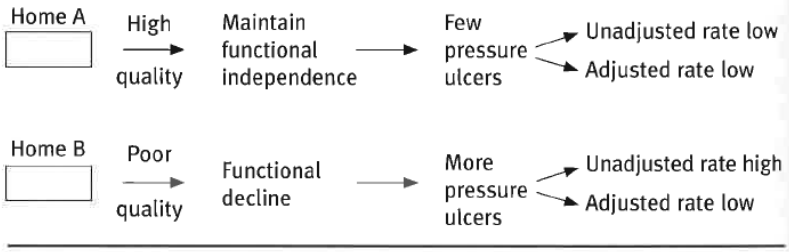
Another decision involves how to scale individual ADL variables in the risk adjustment model. One option is to treat each ADL as a continuous variable. Berlowitz and colleagues (1996a, 2001b) used this approach in developing risk adjustment models for pressure ulcer development in nursing home patients based on bivariable analyses revealing that each 1-point increase in the 5-point mobility and transfer scales produced proportional increases in the rate of pressure ulcer development. Mukamel (1997), in contrast, used a categorical approach in her pressure ulcer model, with dummy variables for each level of the mobility and transfer scales. A third approach is to consider each ADL as dichotomous (dependent/not dependent) and to sum the number of

dependent ADLs for each patient (Porell et al. 1998). None of these approaches for using ADLs is necessarily superior. One must rely on thorough examination of the empirical analyses alongside clinical input when choosing an approach.

Selection of other potential risk factors requires careful thought. Cognitive impairment is common in long-term care and is often reasonable to consider in risk adjustment models. Age is inconsistently associated with outcomes, possibly reflecting the relatively compressed age range typically seen in long-term care settings. Patient-centered information obtained from surveys, while likely to be valuable, generally has not been used for risk adjustment in long-term care. Special considerations in home health care include social support, socioeconomic status, and physical environmental factors. Outcomes of home health care patients may depend more on social support and informal care than on the quality of the home health agency, which suggests that these factors should be considered in performing risk adjustment (Hirdes et al. 2004).

Studies clearly demonstrate that risk adjustment models encompassing multiple dimensions of risk can be developed for long-term care. Nonetheless, two questions arise about the wisdom of this approach. First, many potential risk factors used in models to evaluate providers' performance are themselves products of the quality of care. Inclusion of these risk factors raises the possibility of "over adjusting" or "adjusting out" the quality of care and compromising provider comparisons. Exhibit 16.3 provides an example for the pressure ulcer development outcome. If a nursing home provides poor care, many patients experience declining ADL function. Consequently, many patients also develop a pressure ulcer. The unadjusted rate of pressure ulcer development is high; however, after adjusting for patients' poor functional status, the adjusted rate appears better. This issue is relevant only to long-staying patients whose baseline status for the time interval used in the risk adjustment model probably differs from their status on admission. As is true of hospital care, status at the time of admission does not reflect the quality of nursing home care, so it is an acceptable risk factor. How great a problem this issue is for long-term care risk adjustment remains uncertain. Proposed

EXHIBIT 16.3 Schematic Showing Risks of Overadjustment



solutions include omitting variables linked to the quality of care as predictors and using only admission assessments (Mukamel 1997; Arling et al. 1997).

The second concern is that some risk factors do not necessarily result in poor outcomes when high-quality care is provided (Zimmerman et al. 1995). Because the risk factor does not predict outcomes equally among all providers, difficulties could arise in comparing performance (Zaslavsky 2001).[2] For example, in high-quality nursing homes, the effects of immobility on pressure ulcer development may be mitigated by frequent repositioning. In contrast, due to coexisting neuropathy and vascular disease, diabetes is likely to be similarly important among all patients. Once again, the extent of this concern for long-term care risk adjustment is uncertain.

Time Frame

Consideration of the windows of observation (see Chapter 4) is especially crucial in long-term care, which sees two broad types of patients: those admitted following an acute event who have a high likelihood of recovery and those whose residential location changes and who have little chance of improvement. Patients admitted to long-term care solely for post–acute care need to be evaluated on the basis of quality measures different from those used to evaluate chronic, debilitated patients (Kramer 2002). But even new admissions anticipating lengthy stays are likely to differ from established, long-staying patients in important ways. As a result of their recent illness or surgery, they probably have metabolic derangements, new functional impairments, or side effects from medications. Although these conditions probably have considerable effects on outcomes, long-term care data sets do not adequately capture some sequelae of acute illnesses. Consequently, risk adjustment models have frequently considered "new admission" (or "type of assessment" in the MDS) as a predictor variable to capture elements of disease severity.

Long-staying patients may require intermittent hospitalizations for acute medical events. Capturing these hospitalizations is important; reporting strategies include linking hospital databases or identifying "readmission assessments" in the long-term care database. Often these patients are excluded from the analysis because of uncertainty about whether changes in health status reflect care administered by the hospital or by the long-term care provider.

Time intervals between consecutive records in long-term care databases may vary for each patient. For example, in the MDS, most records involve 90-day intervals (see Exhibit 16.2), but shorter intervals are mandated when the patient's health status significantly changes or on admission for Medicare-paid skilled nursing facility care. Because many outcomes are more likely to appear over longer windows, comparison of outcomes for patients assessed at different intervals (i.e., 90 days versus shorter time

frames) may be inappropriate. Using the MDS to examine pressure ulcer development, one study first defined an index assessment in each calendar quarter and then identified outcomes from as close as possible to 90 days afterward (Berlowitz et al. 2001b). The researchers examined information only from 45 to 135 days following the index evaluation to ensure that the intervals between assessments were similar for all patients.

Observations are "censored" when long-term care patients are discharged or die before the outcomes are assessed (e.g., within the standard 90-day interval for the MDS). This issue frequently arises for new admissions, many of whom stay in nursing homes or receive home care only briefly. Thus, persons with censored observations likely differ in important ways from patients with complete, periodic assessments. The amount of potential bias may depend on the periodicity and timing of the assessments. Exhibit 16.4 shows an example using a VA long-term care database in which an assessment record is produced when patients are admitted and then every April 1 and October 1. The time interval between admission and the next assessment varies from days to up to six months, depending on the date of admission. Patients admitted in February, March, August, or September have two months or less until the next assessment, while for long-staying patients, the interval is exactly six months. Changes in health status for patients admitted and then discharged before the next April 1 or October 1 cannot be determined. Examination of rates of pressure ulcer development found the lowest rate among patients who had less than two months between assessments (Berlowitz et al. 1996b). While functional impairment among patients admitted in February, March, August, or September did not differ from that among patients admitted in other months, patients admitted in those months and remaining for an

EXHIBIT 16.4 Effect of Censoring Patient Assessments on Outcome Rates and Associated Risk Factors*

	Time Between Assessments for Patients Remaining for an Outcome Assessment		
	0 to 2 Months	2 to 4 Months	4 to 6 Months
Number of patients	11,860	5,013	2,994
% developed pressure ulcer	4.0	5.2	5.1
% incontinent	38.6	45.9	44.2
% immobile	54.5	61.1	58.5
% dependent in transfer	28.2	32.0	31.0

*No differences in functional status existed among the groups when residents were examined only at the time of admission. However, Exhibit 16.4 demonstrates significant differences among the residents remaining in long-term care for their outcome assessment. The number of residents also declines sharply over time, indicating the extent of censoring.

April 1 or October 1 outcome assessment were significantly less functionally impaired than were patients who had two to four or four to six months between assessments. These findings likely resulted from selective discharge of healthier patients, after which only "sicker" patients remained. This hypothesis is further suggested by the large decrease in the number of patients as the time between assessments increases, indicating the censoring of many patients. Including the time between assessments among the independent predictors may help address this discrepancy.

Because of extended time frames during which people receive care, some individuals may be included several times in sample model development. For example, MDS assessments are performed every 90 days, so an individual may contribute four records when researchers evaluate care over a year. Because observations for patients included more than once are likely correlated, statistical modeling may need to account for this correlation. Use of multiple observations from the same individual is often preferable to randomly selecting one observation per patient because it provides a larger sample size for modeling. Furthermore, for some outcomes, random selection of one observation falsely elevates the observed rate in the resulting sample.[3]

Examples of Risk Adjustment in Long-Term Care

Performance feedback is central to CMS's current efforts to improve the quality of nursing home care and home care. Different approaches to risk adjustment are being used in long-term care, as suggested by the examples that follow. These methods generally are updated periodically; the CMS website provides the most current information.

Medicare Nursing Home Quality Assessment

Since 2002, the CMS website has publicly displayed nursing home quality measures through the Nursing Home Compare program. Based on MDS data, these quality indicators (QIs) were originally developed by researchers at the Center for Health System Research and Analysis at the University of Wisconsin–Madison (Zimmerman et al. 1995). They were created through a process that combined clinical input, empirical analyses, and field testing. Researchers began with an initial pool of 175 indicators and narrowed it down to 30 items. These QIs were not intended to measure quality directly; rather, they may indicate potential problems that need further investigation by state surveyors or nursing home staff, for example. Thus, they are of limited usefulness to consumers.

These nursing home QIs have undergone further review and refinement (Harris and Clauser 2002). On the basis of conceptual and empirical

reviews, a total of 22 QIs were recommended, including weight loss incidence, pressure ulcer incidence or worsening, deterioration in mood, and prevalence of daily physical restraints (Berg et al. 2002). Ten of these measures were included in the November 2002 release on the Medicare Nursing Home Compare website. Currently, the Nursing Home Compare website reports on 19 measures, of which 14 are related to long-term care residents and 5 are specific to post–acute care. Descriptions of nursing home performance are further simplified for consumers through a five-star rating system that combines quality indicators with information on staffing levels and the results of state surveys.

The US General Accounting Office (2002) (now the US Government Accountability Office) criticized the initial release of Medicare's Nursing Home Compare data on the grounds that it was premature, that accuracy of the MDS remains uncertain, and that the risk adjustment techniques used might be inadequate. Given CMS's goals for the selected QIs and the concerns about overadjustment described earlier, the QIs do not use comprehensive risk adjustment. Instead, they employ a variety of approaches including stratification into high- and low-risk groups, exclusions to limit the denominator, and limited regression models based on up to three risk variables. Research continues to question the limited use of clinically detailed risk adjustment with these measures (Mukamel et al. 2008). One study suggested that case mix explains at least 50 percent of the variation in nursing homes' performance on the urinary/fecal incontinence measure (Li et al. 2010). By failing to address case mix differences, the existing measures are likely biased. How CMS will respond to this research remains unclear.

Medicare Home Care Quality Assessment

CMS has also sponsored efforts to develop risk-adjusted outcome measures for home care using OASIS data (Shaughnessy et al. 2002). As described earlier, OASIS was created to support outcome-based quality improvement and case mix–based reimbursement. Therefore, its quality measures are intended to provide accurate performance data. The researchers first specified outcomes and identified risk factors for each of the outcomes; they then incorporated these data items into the OASIS data collection tool. The 41 outcome measures used in the initial demonstration projects have been incorporated into the CMS OBQI program. Twelve of these measures are reported to consumers on the CMS website as part of Home Care Compare.

The researchers developed risk adjustment models for these outcomes through a detailed process, starting with 149 potential predictors. They used bivariable testing to screen risk factors and then incorporated candidate risk factors into multivariable logistic regression models. They estimated each model several times, revising subsequent models on the basis of clinical input, and evaluated model performance using derivation and validation samples.

The final models include from 20 to over 50 risk factors encompassing multiple dimensions, including functional status, social supports, diagnoses, and care needs. Researchers continue to refine the risk adjustment models by evaluating new predictors and examining alternative statistical techniques (Shaughnessy and Hittle 2002; Nuccio, Goodrich, and Hittle 2008). Models are used to compare home health care agencies' performance to a national reference and to the agencies' own performance in the prior year. Results from the demonstration projects have shown that quality improvement programs incorporating these risk-adjusted data can significantly improve patient outcomes (e.g., reduce hospitalization rates) (Shaughnessy et al. 2002).

Conclusions

Development of risk adjustment strategies for long-term care has progressed considerably. The accuracy and content of existing databases containing detailed clinical information on many long-term care patients are likely to improve, as is the utility of these databases for risk adjustment and quality measurement/improvement. As in other health care settings, information on patient satisfaction and preferences will become increasingly crucial to assessing care. The growing demand for long-term care will also continue to strain the health care system, and care for frail elderly patients will increasingly rely on alternative long-term care settings. Therefore, quality assessment of long-term care and equitable distribution of resources among long-term care settings will become even more critical. Both will require risk adjustment models applicable across the long-term care spectrum. The development of such models will likely challenge the next generation of researchers in long-term care risk adjustment.

Notes

1. Especially in areas experiencing nursing home shortages, concerns arise about "selection bias"—whether some nursing homes specifically avoid patients at potentially high risk of developing problems that could reflect poorly on the facility. In addition, nursing homes that have particular trouble recruiting and retaining even low-level staff may refuse to admit patients who require considerable assistance, even for the most basic activities. All these issues likely relate in complex ways to the quality of nursing homes.
2. Zaslavsky (2001) provides a more detailed description of why it may be problematic when the effect of a given risk factor on the outcome varies by provider. Briefly, consider two providers. For each provider,

the association between disease severity and the outcome rate may be described by a line, "sicker" patients having a higher rate. The lines for the two providers are not parallel when the effect of a risk factor varies. No single value can characterize the difference in performance among providers because it differs depending on how sick patients are.

3. To understand how randomly selecting one observation per patient may falsely elevate the observed rate, consider an outcome, such as development of a pressure ulcer or contracture, that precludes the patient from being included in the sample once it develops. Suppose data on long-staying patients are collected from three consecutive periods. There will be three observations for a patient never developing the condition, one of which will be selected for the sample. In contrast, a patient developing the condition during the second period will have only two observations. Thus, in randomly selecting one observation per patient, the cases that developed the condition are more likely to be selected. As a result, the observed rate in the final sample will be higher than in each of the periods from which observations were selected, which could affect model calibration when the model is applied to individual periods.

CHAPTER

17

THE ROLE OF RISK ADJUSTMENT IN MANAGING HEALTH CARE ORGANIZATIONS

Jennifer Daley

The management of health care organizations becomes more challenging every day as stakeholders demand greater value of health care services and costs escalate as a result of incentives inherent in a fee-for-service delivery system. The aging of the US population, rapid advances in medical technology that extend the length and quality of life, the practice of "defensive medicine," and other factors are also contributing to escalating costs (Bodenheimer 2005a, 2005b; Gawande et al. 2009). *Value* is defined as the quality of care divided by the cost of care (Rosenthal et al. 2007; Chernew et al. 2010; Choudhry, Rosenthal, and Milstein 2010; Porter 2010). Stakeholders seek to maximize this quotient to provide or receive the highest possible quality of care for the lowest possible cost.

The rate of inflation of health care costs outstrips the rate of general inflation by about two to one (PricewaterhouseCoopers 2010). In contrast, the overall quality of care appears to have stayed about the same or, in some cases, declined (McGlynn et al. 2003; Asch et al. 2006; MacDorman and Matthews 2008). Despite dire predictions over the past 30 years that the rate of inflation in health care and US spending on health care would become unsustainable, the United States appears to have reached the point where health care expenditures are threatening the viability of the middle class and the competitiveness of US products and services in the global economy (Banthin and Bernard 2006; Johnson 2012). At the same time, the overall health of the US population ranks far behind other Western countries (Murray and Frenk 2010).

Numerous remedies for the quality/affordability conundrum have been implemented or proposed in the past 30 years. As noted in Chapter 1, in fiscal year 1984 Medicare introduced prospective hospital payment based on Diagnosis-Related Groups (DRGs)—a flat lump sum for inpatient care adjusted by patients' principal diagnosis, surgical procedures, and other diagnoses (complications and comorbidities). This system of acute care hospital payment remains in effect, although the risk adjustment or case classification system has been modified and refined extensively in several iterations, the most recent of which was the 2008 introduction of Medicare Severity DRGs (MS-DRGs).

In the 1990s, many payers introduced capitation, a system under which primary care doctors received a set monthly fee for each patient attributed to them. Primary care physicians were responsible for managing this allocation to ensure it would cover the care of their group or panel of patients over time (Zuvekas and Cohen 2010). Although capitation persists as a form of physician reimbursement in some parts of the United States, it is regarded as a failure for several reasons. First, the risk pools established under capitation were not large enough to withstand an episode of care of very high cost, such as the care of an extremely premature newborn or a trauma patient with extensive injuries. Ultimately, the costs of these very ill patients had to be "carved out" and paid through complicated stop-loss provisions (Maguire et al. 1998). Second, the data systems were insufficiently detailed and not timely enough for physicians and hospitals to manage their patients' care effectively. Third, the risk adjustment methods designed to "level the playing field" for physicians and hospitals that traditionally cared for patients with a higher burden of illness were insufficient to account for the higher utilization and costs of those patients (Iezzoni et al. 1998). Furthermore, patients could change their physicians at will, so when diagnosed with a serious illness, they could switch to a specialist in their disease process, thus creating adverse selection of patients in the risk pools (Post and Carter 1998). The fatal blow to the widespread implementation of capitation was the public's perception that per patient per month reimbursement encouraged underutilization among physicians and restricted patients' access to care (Landon et al. 2004).

As described briefly in Chapter 1, another confluence of events is forcing a major reevaluation of how high-quality, more affordable care is provided. Although the prevalence of people who do not have health insurance coverage ranges widely from Massachusetts (3 percent) to Texas and California (25 percent), approximately 50 million people—one-sixth of the US population—have little or no health insurance coverage (Henry J. Kaiser Family Foundation 2010). As a result of the deep, prolonged recession and slow "jobless" recovery, the number of US citizens who do not have access to health care other than under emergency conditions is increasing and the overall health of the public may deteriorate (AHRQ 2011a). Furthermore, the cost of health insurance continues to increase at a rapid rate. Even after the introduction of health care savings accounts and other financial vehicles that set aside pretax dollars to cover the gap between employer-provided coverage and the actual cost of care, American families are contributing more and more of their wages to health insurance. Given the doubling of the cost of health insurance between 2000 and 2010, American families' income after taxes has effectively remained the same or declined over the past ten years. The cost of providing adequate health care coverage to employees is hindering both global competition for American goods, the small business entrepreneurship that has fostered American innovation and creativity for generations, and

investments in other public programs, such as education, transportation, and rehabilitation of American infrastructure.

Leadership of Health Care Organizations in the Future: The Big Picture

The implications of the quality/affordability conundrum for the management of health care organizations, including providers, payers, and regulators, are profound. Regardless of how payment policy changes under the Affordable Care Act, under risk-based contracting, or in response to employers' demands for lower health insurance premiums, less money per capita will flow into the health care system in the future. As a result, providers' revenue per unit of service (e.g., office visit, procedure or operation, laboratory test, day of home care services) will decrease (Iglehart 2011). Despite lower revenues, expectations for improved quality of care in every dimension will increase. Providers will need to prove to regulators, payers, and the public that they can deliver better quality of care at more affordable costs. Although early expectations about pay for performance and consumer-driven health care have not yet materialized, payers, employers, and regulators are devising methods for incentivizing patients to seek care from lower-cost, higher-quality providers (Bloche 2006). Patients will be expected to pay significantly higher copayments to obtain care from higher-cost, lower-quality providers. The business community is placing intense pressure on commercial payers to lower premiums. Even large, self-insured employers are offering new plans that lower employee contributions to premium costs by requiring enrollees to use limited or restricted networks of providers that provide lower-cost, higher-quality care (Baker et al. 2007).

In addition, the cost of administering health plans is under intense scrutiny; regulations capping administrative costs are being implemented (Jost 2010). Medicare and Medicaid face similar challenges as the number of people covered by these programs continues to increase as a result of the weak economic recovery, the prospect of universal coverage under health care reform (see Chapter 1), and the aging of the baby-boom population (Orszag and Emanuel 2010). Regulators at local, state, and federal levels expect providers to continue to deliver safe, effective care to patients despite declining revenues to providers.

Hospitals and clinicians are accelerating their efforts to improve care while eliminating unnecessary costs and redesigning inefficient systems. Payers are requiring more and more reports about cost and quality from providers. Some reports are used to "steer" patients, while others are made publicly available so consumers may choose high-quality, lower-cost providers. Some payers are withholding a portion of their payments to providers and redistributing this

frontline improvement projects that use small tests of change on a daily or weekly basis to improve processes, but pristine levels of reliability and validity are required for reporting to regulators that subsequently publish comparative quality information, such as mortality rates for specified conditions or surgical procedures at a hospital or physician level regionally or nationally (Krumholz and Normand 2008).

Because health care organizations may function across multiple dimensions of quality (e.g., clinical outcomes, patient satisfaction, safety) and sites of service (e.g., primary care, mental health services, acute hospital-based services), the way in which cost and quality are measured is inherently observational and hence subject to all the biases described elsewhere in this book. Charged with the demands of day-to-day management, organizations rarely have the luxury of designing randomized controlled trials of interventions intended to improve the cost or quality of health care. Any comparison of one institution to another, one provider to another, or one system or department to another (within or across institutions) introduces the possibility of biased results. Sources of bias include differences in patients' health-related risk factors, burden of disease, or random variation resulting from differences in sample size. When presented with new comparative analyses, health care audiences—be they financial managers, clinical leaders and providers, or administrative managers—always pose three key questions:

1. How do we know that the data from which these analyses are derived are reliable, accurate, and a valid measurement of the outcome in question (e.g., cost, quality)? (Often this question takes the form of a protest, i.e., "I don't believe the data.")
2. How do we know that the observed differences in outcomes are not attributable to differences in the populations over which we have no control? (Audiences often express this question by asserting "my patients are sicker.")
3. How do we know that the observed differences are not the result of random variation rather than an intentional intervention on our part? ("I don't think the sample size is big enough to conclude that the results occurred because of the practice or management change I implemented.")

Because health care data are captured or measured over time, the study design of most management reports and analyses is a time series. Sometimes organizations have the luxury of interrupting the time series; they implement an intervention, stop, and then implement it again to study its impact. Often, however, they do not have enough time or an opportunity to conduct such studies. Instead, they generally infer the effectiveness of intentional—and sometimes unintentional—"natural experiments" while caring for patients.

The pace of change has accelerated to the point where organizations have to decide how to change care delivery on the basis of information that is imperfect and of uncertain accuracy and that does not meet the scientifically rigorous standards of a study section review committee or of the editors of a peer-review journal. To offset this problem, they find creative ways to account for those imperfections and uncertainties, often by adapting lessons and techniques from industrial settings to health care.

Statistical Process Control

Unlike managers and leaders in industrial settings, who have control over the input materials and specifications of the products they make, health care managers and leaders have little control over the burden of illness and mix of patients who present at physician offices or hospitals. In general, providers know what outcomes they want to achieve; they also ask patients to identify their goals. Providers strive to make care delivery as efficient and reliable as possible to achieve those goals and outcomes. But health care systems and processes are complex and often opaque to them. Because providers have limited ability to control the processes of care, they rarely, if ever, achieve near-perfect specifications as industrial engineers and manufacturers do. However, industrial methods that have been adapted to the health care setting enable providers to assess the impact they have on processes of care.

Sample sizes are another consideration. Researchers may be able to create robust samples that are clinically meaningful and statistically stable; for clinical laboratory processes, for example, large numbers of tests often are performed within short periods, and therefore researchers may be able to use sampling to study the impact of a process change. Researchers can also estimate the size of an intervention's effect and plan on securing sufficient sample size to avert Type I and Type II errors (see Chapter 12). More often, however, health care organizations have little or no control over the size of the sample of patients who present for care. For example, depending on the size of the hospital, the density of the surrounding population, and the number of hospitals in the same catchment area, it may be difficult to develop a stable point estimate of the median time to thrombolysis in ST segment elevation acute myocardial infarction (AMI).

Statistical process control (SPC) is the application of statistical methods to care processes for purposes of ensuring that care processes maximize patients' clinical outcomes and reduce inefficiencies (Bohmer 2009). Although SPC is used most frequently in manufacturing, it may be applied to any process that has a measurable output. Key SPC tools are control charts, continuous improvement, and small designed experiments, also known as small tests of change. SPC enables analysts to determine whether observed changes in the

output of a process are the result of random variation or caused by the system and system interventions designed to bring about change.

Some variation, called *common cause variation*, is inherent in the design of every system. Variation attributed to intentional modifications to care processes is called *special cause variation*. SPC is perfectly adapted to the ongoing time series experiments used in health care to improve processes and outcomes of care. It provides timely information about the entire process under study and does not rely on retrospective inspection of the process to determine why the desired outcomes did not occur. Retrospective inspection (e.g., root cause analysis of a sentinel event, such as wrong-site surgery) is helpful for determining where the system of care failed (e.g., failure to follow a universal protocol, call a timeout, or use a checklist at the beginning of surgery). SPC is a good way to study whether the process put into place to correct the problem is working as it was intended. SPC also tells analysts when observed changes in a process require no action because the variations are the result of common cause variation.

Exhibit 17.1 shows patients' perceptions of the cleanliness of a hospital. The hospital's leaders were not happy with its cleanliness or with patients' perceptions of that cleanliness as reported in a nationally standardized patient experience survey (Hospital Consumer Assessment of Healthcare Providers

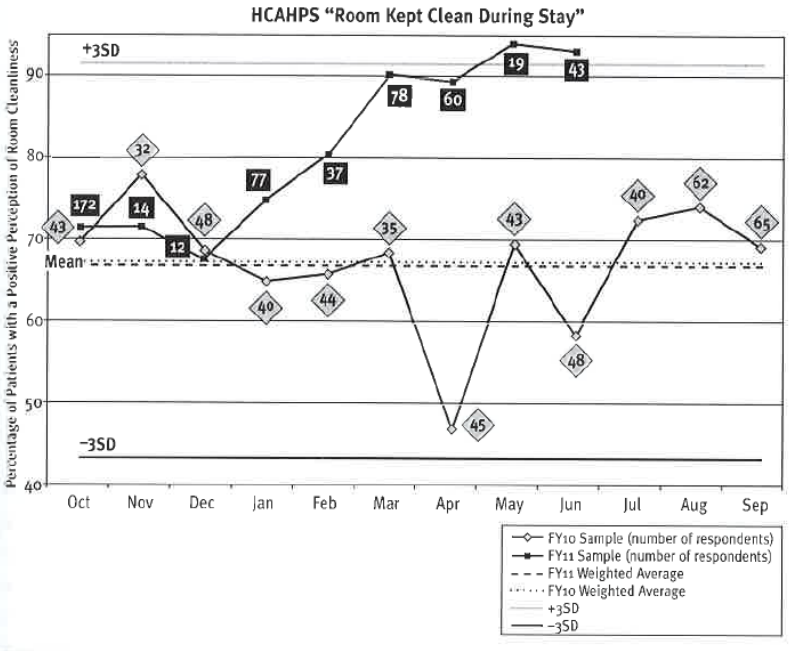


EXHIBIT 17.1 Patients' Perceptions of Hospital Cleanliness Over Time

and Systems [HCAHPS]; see Chapter 7). During the first year of the survey (2010), patients' perceptions of the hospital's cleanliness were below the national HCAHPS average for hospitals of similar size and patient complexity. During one month (April 2010), patients' perceptions dropped significantly below the hospital's own mean value.

In retrospect, this period of especially poor performance coincided with a labor action by the unions representing the hospital's environmental services staff. After the labor action resolved, management and frontline staff collaborated to improve the cleanliness of the facility. The hospital's environmental services staff mapped out the steps involved in cleaning patient rooms using process improvement techniques, such as Ishikawa diagrams, designed experiments, and Pareto charts. They designed small tests of change and quantified the cleaning process. They developed a standard work process, reviewed errors, developed a "buddy system" in which pairs of housekeepers observed each other's work and compared it to the standard process, developed training tools, and used SPC methods and data collection to track the consistency of implementation of these new checklists and working protocols. And most important, they measured patients' perceptions of cleanliness.

Most statistical process charts work best for numeric data with Gaussian distribution assumptions (Alemi, Rom, and Eisenstein 1996). Many statistical rules guide identification of special cause variation, but it is generally defined as

- one point outside the plus/minus 3 sigma limits,
- eight successive consecutive points above (or below) the centerline,
- six or more consecutive points steadily increasing or decreasing,
- two out of three successive points outside of two standard deviations or beyond, or
- fifteen consecutive points outside of one standard deviation on either side of the centerline.

Standard work processes were implemented in October 2010. By early 2011, there was sustained improvement in patients' perceptions of the hospital's cleanliness. For several months at the beginning of calendar year 2011, this improvement was significant (three standard deviations above the prior mean).

Striking a Balance to Achieve Improvement

Hospitals, primary care practices, large multispecialty group practices, and other providers have similar high-level strategic goals. They seek to provide high-quality care from both technical and interpersonal perspectives. They seek to grow, whether organically or through thoughtful investment in specific areas. They need to balance the flow of revenues with associated expenses to achieve a sufficient margin to reinvest in the organization's growth and

resource needs. An engaged workforce of providers and support staff is key to providing patient-centered care, improving patient satisfaction, and securing return visits when patients need care in the future. Service to the community may also be an important part of overall strategy. Hypothetically, a provider organization can sacrifice one of these strategic goals to the others, but in today's rapidly changing health care environment, ignoring one of these strategic goals to strengthen the others is ill-advised.

In health care, these strategic goals—quality and safety, customer experience, employee engagement, fiscal stewardship, growth, and community service—form the "pillars" of a balanced scorecard of measures, a tool used to assess the overall stability and success of an organization (Studer 2009). Exhibit 17.2 is an example of a balanced scorecard for an acute care hospital. This high-level set of measures might inform the senior clinical and administrative leaders of the hospital or be the basis of regular reporting to the hospital's governance. Periodically (e.g., annually), the clinical and administrative leaders of the hospital choose areas of focus from each pillar. For example, in quality and safety, they may balance measures of patient safety (e.g., reducing central-line infections and medication errors) with improving the processes of care for key clinical conditions (e.g., AMI, pneumonia, and heart failure). Increasingly, under value-based purchasing (see Chapter 1), some reimbursement and public reporting depend on demonstrating high-quality care and reducing adverse events, which are measured by external national standards and metrics endorsed by the Centers for Medicare & Medicaid Services (CMS), The Joint Commission, or other accrediting bodies.

In addition to measures and improvement initiatives driven by external stakeholders, the hospital may have internal improvement initiatives (e.g., reducing mortality from sepsis by instituting a rapid response sepsis team). Those measures may appear on the balanced scorecard as well. For example, the leaders of the hospital in Exhibit 17.2 chose to focus on reducing central-line infections, reducing patient falls with injury, and improving median time to coronary artery opening for ST segment elevation AMI. Long waiting times for new and return appointments in the affiliated primary care practices had driven patient dissatisfaction, and the hospital leaders initiated an improvement project to increase patient access to primary care appointments. In concert with the governing board, the hospital leaders recognized that providing patient-centered care to the community's patients and their families, responding to the needs of the medical staff, and contributing to the health and well-being of the community were their primary service goals; the metrics in the service pillar reflect those priorities. The leaders also realized that having an engaged, stable workforce that embraces the vision and mission of the hospital and community is a key factor in patients' and physicians' experiences. Metrics reflecting the results of an employee engagement survey and the rate of employee turnover provided the basis for a high-level assessment

EXHIBIT 17.2 Acute Care Hospital Balanced Scorecard: Dimensions, Example Measures, and Adjustment for Case Mix and Severity of Illness

Dimension	Goal	Example Measure	Data Source	Adjustment for Case Mix and Severity of Illness
Service	Improve patient experience	HCAHPS—nurse communication	Patient survey	• Stratified by emergency room, surgery, obstetrics, medicine • Stratified by inpatient care unit
	Improve physician satisfaction	Physician satisfaction survey—support service responsiveness (e.g., laboratory, imaging)	Physician survey	• Stratified by type of physician (e.g., primary care physician, hospitalist, general surgeon, intensivist) • Stratified by age of physician, specialty, volume of admissions/visits to the hospital
	Provide service to the community the hospital serves	Percentage of patient care revenues devoted to community service	General ledger and hospital operations	None
Quality	Improve patient safety: reduce venous catheter-associated bloodstream infections	Central venous catheter bloodstream infections (CVC-BSI)/1,000 catheter days	Infection control monitoring database and electronic medical record	Stratified by those identified in adult intensive care unit (ICU), neonatal ICU, or acute adult inpatient service
	Improve patient safety: reduce patient falls with injury	Patient falls with injury/1,000 patient days	Nursing and/or patient safety/risk management system	Stratified by type of acute care unit (e.g., adult medical, adult surgical, patient care unit)
	Improve evidence-based practice: acute myocardial infarction	Door-to-balloon time for ST segment elevation myocardial infarction (STEMI)	CMS core measures	• Applies only to STEMI patients • Stratification of denominator by distance from the hospital at time of onset of symptoms
	Improve patient access to care: reduce waiting times for elective primary care appointments	Time to first available "new patient" appointment in primary care practice	Appointment scheduling system	Urgent/emergent/routine new patient visit

EXHIBIT 17.2 (Continued)

Dimension	Goal	Example Measure	Data Source	Adjustment for Case Mix and Severity of Illness
Financial stewardship	Reduce costs	Cost per case mix–adjusted discharge	Financial management systems	Cost adjusted by case mix and severity assignments in DRG algorithms
	Improve employee productivity	Work hours/ unit of service	Payroll and human resources database	• Stratified by type of employee (e.g., registered nurse, housekeeper, pharmacist) • Stratified by acute-care unit (e.g., ICU, outpatient service)
	Improve overall financial performance	Operating margin	General ledger	None
	Achieve budget revenues and expenses	Actual revenue and expenses as a ratio of budgeted expenses	Budget system, cost accounting system, general ledger	• May be stratified by service line (e.g., cardiology, general surgery) or diagnosis/ procedure • May be stratified by DRGs or groups of clinically related DRGs
People	Retain excellent, committed workforce	Employee turnover—percentage of employees leaving the organization by choice	Human resources database	• May be stratified by unit, manager, or senior management • May be stratified by employment tenure
	Promote employee engagement in the mission and performance of the organization	Employee engagement survey: "My supervisor [or someone at work] cares about me as a person."	Employee engagement survey	• May be stratified by unit, manager, or senior management • May be stratified by employment tenure
	Reduce use of temporary help	Hours of overtime and per diem expenditures	Human resources database and payroll	Stratify by human resources cost center, manager, and senior manager

(Continued on next page)

EXHIBIT 17.2 (Continued)

Dimension	Goal	Example Measure	Data Source	Adjustment for Case Mix and Severity of Illness
Growth	Increase volume	Volume of surgery	Hospital administrative systems	Case mix adjusted using DRGs or related DRGs
	Increase volume of new patients coming to the hospital	New patient covered lives	Managed care system	None
	Growth in service line targeted by growth strategy	Increase in volume and revenue; increase in combined hospital and physician group margin	Hospital and practice management systems	Case mix adjusted by DRGs or related DRGs; adjusted for episode of care with grouping methodology; revenue by relative value work unit

of employees' satisfaction with the hospital workplace. Because the cost of labor is a predominant driver of the cost of care (compensation and benefits account for 55–60 percent of a hospital's operating costs), the hospital also aimed to reduce the use of "premium" labor, specifically overtime by staff and employment of per diem workers.

Maintaining the fiscal health of the hospital and generating resources to invest in the workforce, infrastructure, and information technology needs of the hospital are also paramount. Likewise, strategic areas in which opportunities to grow services needed by the community are identified, and detailed plans to build and market those services are developed. The dimensions of financial stewardship and growth are also included in Exhibit 17.2. Systems and processes of care that maximize the use of employees' time and skills are vital. By maintaining an excess of operating revenue over operating expenses—known as *margin*—the hospital is able to meet demands from the community, patients, and medical staff for improvements to the physical plant, improved equipment, and improved information technology.

Margins will become increasingly difficult to maintain as revenues decrease but the demands for high-quality patient-centered services, infrastructure, and technology increase. A decade ago, a thriving community hospital or academic health center could expect to maintain an operating margin of 5–7 percent. Today, operating margins of 3 percent are considered excellent, and a significant minority of hospitals have negative margins. Unless other sources of revenue (e.g., philanthropy, investment income) subsidize their operating margins, those hospitals face an uncertain future. Some will close; others will be acquired by other hospitals or hospital systems.

The Role of Risk Adjustment in Comparing Performance Across Health Care Organizations

Leaders of health care organizations often choose metrics for their performance scorecard that are collected in the same or similar ways by different hospitals. When those data are available, one organization's performance can be compared to that of similar institutions, competitors, and organizations. Comparison across hospitals, physician practices, and other health care institutions immediately raises issues of risk adjustment. Valid comparisons show providers areas in which they can improve and achieve better results and help them determine whether the processes of better-performing organizations can be translated to their own.

As previously noted, a key first step is to present performance measures based on accurate and reliable data with good face validity to clinical and administrative leaders. The next step is to reassure them that the "playing field is even" or that the presentation "compares apples to apples." With rare exceptions, "my patients are sicker" is clinicians' and administrators' response when they see comparative data indicating they or their organization do not perform as well as they had expected. To make valid comparisons in health care management, analysts primarily use two methods of risk adjustment: stratification and multivariate methods.

Stratification—the division of a population into subgroups that are similar in some way—is a common approach to risk adjustment (see Chapter 12). The DRG methods, episode grouping methodologies (see Chapter 4), and ambulatory groupers all stratify patients or encounters (e.g., admissions, office visits, a year of the care of a patient with a chronic illness) by using inclusion and exclusion criteria to identify a specific denominator of patients. Exhibit 17.2 identifies some of the prevalent methods used to risk-adjust the metrics included in a hospital's balanced scorecard. Inclusion criteria may encompass patients who have a specific diagnosis or who have undergone a certain procedure. To make the patient population as homogeneous as possible, analysts exclude patients from a cohort if clinically they are more appropriate for inclusion in another cohort.

Exhibit 17.3 lists the inclusion and exclusion criteria for several of the core measures used by CMS, The Joint Commission, and other organizations to assess quality of care. The numerators in these process measures are relatively easy to determine. For example, for the influenza vaccination core measure (the third measure listed in Exhibit 17.3), a patient either did or did not receive the influenza vaccine. The denominator would be determined by establishing which hospitalized patients should be in the cohort of persons eligible to receive the vaccine through careful screening for the exclusion criteria (e.g., the patient has a hypersensitivity to eggs and cannot receive the

EXHIBIT 17.3 Risk Adjustment of Process Measures by Stratification: CMS Core Measure Inclusion and Exclusion Criteria

Core Measure	Inclusion Criteria for Denominator	Exclusion Criteria for Denominator	Numerator
ST segment elevation myocardial infarction (STEMI); median time to primary thrombolysis	Discharges with an ICD-9-CM principal diagnosis code for AMI as defined in Appendix A, Table 1.1* **and** percutaneous coronary intervention (PCI) (ICD-9-CM principal and other procedure codes for PCI as defined in Appendix A, Table 1.2)* **and** ST segment elevation or left bundle branch block (LBBB) on the electrocardiogram (ECG) performed closest to hospital arrival **and** PCI performed within 24 hours after hospital arrival	• Patients younger than age 18 • Patients who have a length of stay (LOS) greater than 120 days • Patients enrolled in clinical trials • Patients received as a transfer from an inpatient or outpatient department of another hospital • Patients received as a transfer from the emergency/observation department of another hospital • Patients received as a transfer from an ambulatory surgery center • Patients administered a fibrinolytic agent prior to PCI • PCI described as non-primary by a physician/advanced practice nurse (APN)/physician assistant (PA) • Patients who did not receive PCI within 90 minutes of arrival at the hospital and had a reason for delay documented by a physician/APN/PA (e.g., social, religious, initial concern or refusal, cardiopulmonary arrest, balloon pump insertion, respiratory failure requiring intubation)	Time (in minutes) from hospital arrival to primary PCI for patients with ST segment elevation or LBBB on the ECG performed closest to hospital arrival

EXHIBIT 17.3 (Continued)

Core Measure	Inclusion Criteria for Denominator	Exclusion Criteria for Denominator	Numerator
Congestive heart failure: angiotensin converting enzyme inhibitor (ACEI) or angiotensin receptor blocker (ARB) for left ventricular systolic dysfunction	Discharges with an ICD-9-CM principal diagnosis code for heart failure as defined in Appendix A, Table 2.1* **and** chart documentation of a left ventricular ejection fraction less than 40% or a narrative description of left ventricular systolic function consistent with moderate or severe systolic dysfunction	• Patients who had a left ventricular assistive device (LVAD) or heart transplant procedure during hospital stay (ICD-9-CM procedure code for LVAD and heart transplant as defined in Appendix A, Table 2.2)* • Patients younger than age 18 • Patients who have an LOS greater than 120 days • Patients enrolled in clinical trials • Patients discharged to another hospital • Patients who left against medical advice • Patients who expired • Patients discharged to home for hospice care • Patients discharged to a health care facility for hospice care • Patients with "comfort measures only" documented • Patients with a documented reason for no ACEI and no ARB at discharge	Heart failure patients who are prescribed an ACEI or ARB at hospital discharge

(Continued on next page)

EXHIBIT 17.3
(Continued)

Core Measure	Inclusion Criteria for Denominator	Exclusion Criteria for Denominator	Numerator
Influenza vaccination	• Patients who received the influenza vaccine during current inpatient hospitalization • Patients who have an ICD-9-CM principal procedure code or other procedure codes from Table 12.9 for prophylactic vaccination against influenza during this inpatient hospitalization* • Patients who received the influenza vaccine during the current year's flu season but prior to the current hospitalization • Patients who were offered and declined the influenza vaccine • Patients who have an allergy/sensitivity to the vaccine or for whom the vaccine is not likely to be effective due to the following: —Hypersensitivity to eggs or other component(s) of the vaccine —History of Guillain-Barré syndrome within six weeks after a previous influenza vaccination —Bone marrow transplant within the past six months —Anaphylactic latex allergy	None	Inpatient discharges who were screened for influenza vaccine status and were vaccinated prior to discharge if indicated

EXHIBIT 17.3 (Continued)

Core Measure	Inclusion Criteria for Denominator	Exclusion Criteria for Denominator	Numerator
Venous thromboembolism (VTE) prophylaxis	All patients	• Patients younger than age 18 • Patients who have an LOS less than 2 days and greater than 120 days • Patients with "comfort measures only" documented on day of or day after hospital arrival • Patients enrolled in clinical trials • Patients who are direct admissions to intensive care unit (ICU) or transferred to ICU on day of or day after hospital admission who have an ICU LOS greater than or equal to one day • Patients with ICD-9-CM principal diagnosis code of mental disorders or stroke as defined in Appendix A, Table 7.01, 8.1, or 8.2* • Patients with ICD-9-CM principal or other diagnosis codes of obstetrics or VTE as defined in Appendix A, Table 7.02, 7.03, or 7.04* • Patients with ICD-9-CM principal procedure code of Surgical Care Improvement Project VTE selected surgeries as defined in Appendix A, Table 5.17, 5.19, 5.20, 5.21, 5.22, 5.23, or 5.24*	Patients who received VTE prophylaxis or have documentation why no VTE prophylaxis was given • on the day of or day after hospital admission, or • on the day of or day after surgery end date for surgeries that start the day of or day after hospital admission

*See www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1228767363466.

Source: CMS (2011c).

vaccine). For the STEMI measure (the first measure listed in Exhibit 17.3), patients entering the hospital with an ST segment elevation AMI who are terminally ill and receiving comfort measures only are not eligible for percutaneous angioplasty and therefore should be excluded from the denominator.

Stratification is commonly used to ensure relatively homogeneous populations so that comparisons of process or outcome measures are more equitable for providers. Nonetheless, stratification has limitations. If numerous patient characteristics are required to risk-adjust adequately, stratification can lead to very small numbers of patients in the strata and hinder meaningful analysis. Stratification also does not allow weighting of the risk variables.

When multiple attributes of patients may bias comparisons, multivariable analysis is the most appropriate technique for risk adjustment (see Chapter 10). For example, analysts may use multivariable analysis when working with mortality data reflecting outcomes that could be influenced by multiple patient characteristics. The resources available at different institutions may also influence patient outcomes. In these circumstances, characteristics of a cohort of patients, coded in administrative data or identified through chart review, are analyzed. The relative weight of the individual characteristics can be determined using multivariable analysis, and a predicted probability of the outcome is generated for each patient in the cohort and compared to the actual outcome. These analyses can also be done in a "hierarchical" way so that hospital characteristics as well as patient characteristics are incorporated into the analysis (Chapter 12).

As described in chapters 3, 9, and 12, CMS publishes risk-adjusted 30-day mortality rates using Medicare Part A claims data for AMI, pneumonia, and congestive heart failure. Administrative data have significant limitations because important clinical variables associated with a high risk of mortality may not coincide with a precise ICD-9-CM code (see Chapter 5). By supplementing the administrative data with data from the Social Security Administration, CMS can follow the patients for the entire 30-day post-admission period and confirm the patients' vital status at 30 days after admission. Exhibit 17.4 shows how the cohorts are identified by ICD-9-CM codes using inclusion and exclusion criteria; it also lists patients' cardiovascular and other comorbid conditions present on admission that may be associated with higher 30-day mortality rates.

Conclusion

An exhaustive description of all the metrics used to measure the performance of health care organizations is beyond the scope of this book. The dimensions of quality, affordability, and performance vary by patient population, illness, treatment, and site of care (e.g., home care, skilled nursing, primary

EXHIBIT 17.4 CMS Hospital Compare Risk Adjustment of Hospital 30-Day Mortality Rates

30-Day Hospital Mortality Rate	Inclusion Criteria	Exclusion Criteria	Risk Adjustment Independent Variables*		Statistical Methodology
Acute myocardial infarction (AMI): hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following AMI hospitalization	• Admissions for Medicare fee-for-service (FFS) and Veterans Administration (VA) beneficiaries aged 65 or over discharged from nonfederal acute care hospitals or VA hospitals, respectively, having a principal discharge diagnosis of AMI • CMS FFS beneficiaries who have an index hospitalization in a nonfederal hospital are included if they have been enrolled in Medicare Part A and Part B for the 12 months prior to the date of admission to ensure a full year of administrative data for risk adjustment. (This requirement does not apply to patients who have an index admission in a VA hospital.) • For patients who have more than one admission in a given year for a given condition, only one admission is randomly selected to include in the cohort (the others are excluded). An index admission is the hospitalization considered for mortality outcome determination. • The measure includes patients who are admitted to a VA or nonfederal acute care hospital with a diagnosis of AMI and then transferred to another acute care facility (VA or nonfederal) if the primary discharge diagnosis is AMI at the second hospital. The hospital from which a patient is transferred is designated as the institution responsible for the outcome.	• Patients who were discharged on the day of admission or the following day and did not die or were not transferred (because it is less likely they had a diagnosis of AMI) • Patients with an inconsistent or unknown mortality status or other unreliable data (e.g., date of death precedes admission date) • Patients transferred from another acute care hospital or VA hospital (because the death is attributed to the hospital to which the patient was initially admitted) • Patients enrolled in the Medicare hospice program any time in the 12 months prior to the index hospitalization, including the first day of the index admission (because it is likely these patients are continuing to seek comfort measures only) • Patients discharged against medical advice (AMA) (because providers did not have the opportunity to deliver full care and prepare the patient for discharge) • Admissions that were not the first hospitalization in the 30 days prior to a patient's death**	Demographics	• Age 65 (years over 65, continuous) • Male	Hierarchical logistic regression modeling
			Cardiovascular condition	• History of percutaneous transluminal coronary angioplasty (PTCA) • History of coronary artery bypass graft (CABG) • Congestive heart failure • History of AMI • Unstable angina • Anterior myocardial infarction • Other location of myocardial infarction • Chronic atherosclerosis • Cardiorespiratory failure and shock • Valvular and rheumatic heart disease	
			Comorbidity	• Hypertension • Stroke • Cerebrovascular disease • Renal failure • Chronic obstructive pulmonary disease • Pneumonia • Diabetes and complications of diabetes mellitus • Protein-calorie malnutrition • Dementia and senility • Hemiplegia, paraplegia, paralysis, and functional disability • Peripheral vascular disease • Metastatic cancer, acute leukemia, and other severe cancers • Trauma in the last year • Major psychiatric disorders • Chronic liver disease	

(Continued on next page)

EXHIBIT 17.4 (Continued)

30-Day Hospital Mortality Rate	Inclusion Criteria	Exclusion Criteria	Risk Adjustment Independent Variables*		Statistical Methodology
Community-acquired pneumonia: hospital 30-day, all-cause, RSMR following pneumonia hospitalization	• Admissions for Medicare FFS and VA beneficiaries aged 65 or over discharged from nonfederal acute care hospitals or VA hospitals, respectively, who have a principal discharge diagnosis of pneumonia • CMS FFS beneficiaries who have an index hospitalization in a nonfederal hospital are included if they have been enrolled in Medicare Part A and Part B for the 12 months prior to the date of admission to ensure a full year of administrative data for risk adjustment. (This requirement does not apply to patients who have an index admission in a VA hospital.) • For patients who have more than one admission in a given year for a given condition, only one admission is randomly selected to include in the cohort (the others are excluded). An index admission is the hospitalization considered for mortality outcome determination. • The measure includes patients who are admitted to a VA or nonfederal acute care hospital with a diagnosis of pneumonia and then transferred to another acute care facility (VA or nonfederal) if the primary discharge diagnosis is pneumonia at the second hospital. The hospital from which a patient is transferred is designated as the institution responsible for the outcome.	• Patients who were discharged on the day of admission or the following day and did not die or were not transferred (because it is less likely they had a diagnosis of pneumonia) • Patients with an inconsistent or unknown mortality status or other unreliable data (e.g., date of death precedes admission date) • Patients transferred from another acute care hospital or VA hospital (because the death is attributed to the hospital to which the patient was initially admitted) • Patients enrolled in the Medicare hospice program any time in the 12 months prior to the index hospitalization, including the first day of the index admission (because it is likely these patients are continuing to seek comfort measures only) • Patients discharged AMA (because providers did not have the opportunity to deliver full care and prepare the patient for discharge) • Admissions that were not the first hospitalization in the 30 days prior to a patient's death**	Demographics	• Age 65 (years over 65, continuous) • Male	Hierarchical logistic regression modeling
			Cardiovascular condition	• History of PTCA • History of CABG • Congestive heart failure • AMI • Unstable angina • Chronic atherosclerosis • Cardiorespiratory failure and shock	
			Comorbidity	• Hypertension • Stroke • Cerebrovascular disease • Renal failure • Chronic obstructive pulmonary disease • Pneumonia • Protein-calorie malnutrition • Dementia and senility • Hemiplegia, paraplegia, paralysis, and functional disability • Peripheral vascular disease • Metastatic cancer, acute leukemia, and other severe cancers • Trauma in the last year • Major psychiatric disorders • Chronic liver disease • Severe hematological disorders • Iron deficiency/anemias/blood disease • Depression • Parkinson's/ Huntington's diseases • Seizure disorders and convulsions • Fibrosis of lung and other chronic lung disorders • Asthma • Vertebral fractures	

EXHIBIT 17.4 (Continued)

30-Day Hospital Mortality Rate	Inclusion Criteria	Exclusion Criteria	Risk Adjustment Independent Variables*		Statistical Methodology
Congestive heart failure: hospital 30-day, all-cause, RSMR following heart failure (HF) hospitalization	• Admissions for Medicare FFS and VA beneficiaries aged 65 or over discharged from nonfederal acute care hospitals or VA hospitals, respectively, who have a principal discharge diagnosis of HF • CMS FFS beneficiaries who have an index hospitalization in a nonfederal hospital are included if they have been enrolled in Medicare Part A and Part B for the 12 months prior to the date of admission to ensure a full year of administrative data for risk adjustment. (This requirement does not apply to patients who have an index admission in a VA hospital.) • For patients who have more than one admission in a given year for a given condition, only one admission is randomly selected to include in the cohort (the others are excluded). An index admission is the hospitalization considered for mortality outcome determination. • The measure includes patients who are admitted to a VA or nonfederal acute care hospital with a diagnosis of HF and then transferred to another acute care facility (VA or nonfederal) if the primary discharge diagnosis is HF at the second hospital. The hospital from which a patient is transferred is designated as the institution responsible for the outcome.	• Patients who were discharged on the day of admission or the following day and did not die or were not transferred (because it is less likely they had a diagnosis of HF) • Patients with an inconsistent or unknown mortality status or other unreliable data (e.g., date of death precedes admission date) • Patients transferred from another acute care hospital or VA hospital (because the death is attributed to the hospital to which the patient was initially admitted) • Patients enrolled in the Medicare hospice program any time in the 12 months prior to the index hospitalization, including the first day of the index admission (because it is likely these patients are continuing to seek comfort measures only) • Patients discharged AMA (because providers did not have the opportunity to deliver full care and prepare the patient for discharge) • Admissions that were not the first hospitalization in the 30 days prior to a patient's death**	Demographics	• Age 65 (years over 65, continuous) • Male	Hierarchical logistic regression modeling
			Cardiovascular condition	• History of PTCA • History of CABG • Congestive heart failure • AMI • Unstable angina • Chronic atherosclerosis • Cardiorespiratory failure and shock • Valvular and rheumatic heart disease	
			Comorbidity	• Hypertension • Stroke • Renal failure • Chronic obstructive pulmonary disease • Pneumonia • Diabetes and complications of diabetes mellitus • Protein-calorie malnutrition • Dementia and senility • Hemiplegia, paraplegia, paralysis, and functional disability • Peripheral vascular disease • Metastatic cancer, acute leukemia, and other severe cancers • Trauma in the last year • Major psychiatric disorders • Chronic liver disease	

*The final set of risk adjustment variables included demographics, cardiovascular condition, and comorbidity.

**This exclusion criterion is applied after one admission per patient per year is randomly selected. It applies only when two randomly selected admissions occur during the transition months (December and January for calendar year data) and the patient subsequently dies. For example: A patient is admitted on December 18, 2006, and readmitted on January 2, 2007; the patient dies on January 15, 2007. If both of these admissions were randomly selected for inclusion (one for the 2006 calendar year period and the other for the 2007 calendar year period), the January 2, 2007, admission would be excluded so that the death would not be assigned to two admissions (one in 2006 and one in 2007).

Source: CMS (2011c).

care ambulatory office practice, acute care hospital). They also vary by the perspective of those monitoring performance, such as providers, health plans, and regulators. Universally, however, questions about the reliability and validity of the data sources used for comparison will arise. In addition, any comparison between groups of patients or providers from observational data will raise concerns about whether factors out of the control of patients or providers may have confounded the comparisons. These objections typically arise from providers whose performance is lower than other providers (i.e., "But my patients are sicker!"). Most providers are convinced that adequate risk adjustment will eliminate any observed variation in quality and affordability. In some cases, the observed variation is eliminated by risk adjustment. More challenging, when significant variation persists after risk adjustment, an opportunity is presented to study the process of care to identify meaningful ways of improving patient care—the most important outcome of performance measurement.

CHAPTER

18

FINAL OBSERVATIONS

Lisa I. Iezzoni

The four editions of this book span nearly two decades. In the American health care system, this expanse seems like an eternity. As we prepared the first edition, President Bill Clinton's health care reform proposals preoccupied public discourse with bold assertions about protecting the best in American medicine (White House Domestic Policy Council 1993, 100):

> If we reformed everything else in health care, but failed to preserve and enhance the high quality of our medical care, we will have taken a step backward, not forward. Quality is something that we simply can't leave to chance.

At the time, report cards on provider performance aimed to inform consumers and purchasers alike. As Dennis S. O'Leary (1993, 487), then president of The Joint Commission on Accreditation of Healthcare Organizations (now called simply The Joint Commission), stated: "Report card day is coming in the health care world." Report cards were intended to help individuals and organizations make informed choices and facilitate oversight of health care quality as President Clinton's proposed "managed competition" took hold. Of course, Clinton's health care reform package imploded spectacularly within the year (Blumenthal and Morone 2009).

Three years later, as we completed the second edition, managed care was ascendant. Countless divergent regional initiatives had balkanized health care reform (Iglehart 1994). Performance profiling was aimed at facilitating vigorous competition among both individual providers and managed care plans. Providers and plans thus battled on the basis of value, an amalgam of cost and quality (see Chapter 17). The public suspected that the emphasis on performance profiling was merely a ruse, believing the much maligned managed care companies cared only about controlling costs. Managed care plans used performance profiles to inform contracting decisions and to target individual physicians for in-depth review (Goldfield and Boland 1996). For New York State coronary artery bypass graft (CABG) surgeons, public reporting of individual performance data became routine, and other clinicians anticipated (feared) they would be next. As Blumenthal and Epstein (1996, 1329) observed:

> It seems fair to say that in the area of quality measurement and reporting, physicians can expect little relief from the feeling that they increasingly work in a fishbowl and are being judged by groups and measures with which they have little familiarity. Managing this reality is one of the greatest challenges confronting the profession at the current time.

As we finished the third edition, managed care was retreating, sidelined largely by consumers' persistent demands for options and choices. Costs again were rising dramatically, the number of uninsured persons was still increasing, several authoritative Institute of Medicine (IOM 2001a; IOM Committee on Enhancing Federal Healthcare Quality Programs 2002) reports raised pointed and troubling questions about the quality of US health care, and some politicians cautiously revisited long-dormant proposals for universal coverage. Meanwhile, *pay for performance* became the prevailing mantra. This phrase had a hard-headed ring that the perhaps euphemistic *report card* and *performance profiling* did not: Its intention was crystal clear.

As this fourth edition nears completion, uncertainty reigns. As noted in Chapter 1, disgruntled politicians, especially those associated with the tea party, are assaulting the Affordable Care Act (ACA), and the US Supreme Court is scrutinizing its constitutionality. In addition, as Rosenthal (2011, 1887) notes, "The US national debt looms like a black cloud on the horizon, and Medicare and Medicaid, which account for about 20% of the federal budget, are responsible for a disproportionate share of projected growth in federal spending." White House proposals for controlling unsustainable cost growth include refocused pay-for-performance schemes, accountable care organizations (ACOs), and episode-based payment. How to operationalize ACOs is a major question posed by all sides (Ginsburg 2011; Rosenthal, Cutler, and Feder 2011). Furthermore, according to Rosenthal (2011, 1889):

> Focusing reform on provider payment changes also carries risks. Performance-based or global payments may encourage providers to avoid high-risk patients or game the system (e.g., by unbundling services from episode-based payments), although risk adjustment and auditing can minimize these risks. Providers may also not be ready to assume the clinical management and financial risks required to accept the kinds of payment arrangements envisioned in the President's plan.

Solutions to these enormous challenges are not yet in sight.

Risk Adjustment as a Common Thread

Risk adjustment links the four defining concepts from these recent mini-epochs in the US health care system—the use of report cards to support managed competition, performance profiling to inform "unmanaged" competition, pay for performance or value-based purchasing, and episode-based payment to control intractably rising costs. In operational terms, these four concepts involve similar approaches. All demand comparisons of patients' outcomes across clinicians, hospitals and other facilities, health plans, or health care delivery systems, and meaningful comparisons generally require risk adjustment.

As noted throughout this book, risk adjustment is relatively well developed for certain purposes, such as for setting prospective payment rates

and evaluating outcomes of specific interventions (e.g., ICU care and CABG surgery). Using even imperfect risk adjustment methods for internal quality improvement activities is well accepted and has proven remarkably effective for some hospitals and health care delivery systems. However, clinicians may hesitate to change or critically examine their practices when shown information they see as flawed. Hospitals and other provider groups therefore need risk-adjusted information to promote productive dialogue with their clinical staffs. As noted in Chapter 17, once clinicians believe the information, they will discuss ways to change their practices to improve the quality and efficiency—the value—of care.

Problems may arise when comparisons of risk-adjusted data move beyond the safe confines of individual practice settings to the external world. For example, as noted in Chapter 2, clinicians and purchasers may define even basic terms differently. Drawing on points raised in Chapter 17, Exhibit 18.1 summarizes almost diametrically opposed common views held by many clinicians and purchasers. At the bottom line, clinicians worry about being treated unfairly or punitively on the basis of data they view as fundamentally flawed. In contrast, purchasers want providers to assume responsibility for the cost and quality of their services, and they argue against waiting for a "perfect" risk-adjustment methodology, which will probably never materialize.

The rhetoric of today's debates about performance measures sometimes evokes a memory of Ernest Amory Codman (1934, xxiii) from the early twentieth century (see Chapter 1): "Comparisons are odious, but comparison is necessary in science. Until we freely make therapeutic comparisons, we cannot claim that a given hospital is efficient, for efficiency implies that the results have been looked into." The difference today is the huge size of the stakes. Although concerns about the fairness of risk-adjusted prospective payment remain, this area is often treated as an arcane issue amenable to technical patches (e.g., partial capitation, special payments for outliers). The more heated exchanges involve using risk-adjusted outcomes information to draw inferences about the quality of care and to implement value-based purchasing.

Little objective, independent information is available about the meaning of risk-adjusted outcomes. Do worse risk-adjusted outcomes truly reflect worse quality of care? This lack of information suggests several important considerations for initiatives that, for example, use risk-adjusted information as a quality indicator when creating pay-for-performance programs or publishing performance profiles on the Internet.

First, participants should acknowledge that they are jointly conducting an applied experiment. In an experiment—as opposed to endeavors that use well-accepted, rigorous methods and produce clearly understood benefits—evaluation is critical. Some initiatives have recognized this concern. For example, one of the earliest initiatives to publicly release mortality rates for specific institutions (and soon for individual surgeons) was New York State's

EXHIBIT 18.1 Common Concerns of Clinicians and Purchasers

Clinicians

- The risk adjustment method is not clinically credible; therefore, comparing risk-adjusted outcomes across clinicians is neither valid nor meaningful.
- The risk-adjusted data focus on limited outcomes (e.g., in-hospital or 30-day mortality for specific conditions) and do not address quality concerns most relevant to patients and clinicians.
- Those examining the risk-adjusted data do not understand the limitations of the risk adjustment methods, the data, or the analytic techniques and therefore draw inappropriate conclusions and make unwarranted decisions.
- Collection of clinical information for risk adjustment is essential for clinical credibility but is logistically burdensome and adds another administrative expense to an already stressed system.
- The data do not provide useful, specific information about how to change care to improve outcomes.
- Imperfect data are used wrongly to determine payment levels and to threaten clinicians (e.g., withdraw health plan contracts, put providers into insurance tiers requiring higher copayments from health plan enrollees).
- When risk-adjusted data are widely reported on the Internet or by the media, few caveats are presented about their limitations; the general public therefore receives erroneous impressions about provider performance, and reputations suffer.

Purchasers

- Given unsustainable increases in health care costs, providers must be held accountable for what they produce.
- We must align incentives to produce the best quality care per dollar spent; multiple IOM and other reports document persistent quality shortfalls and medical errors.
- Purchasers and consumers require information to make informed decisions about where to obtain health care.
- Some risk-adjusted outcomes measures are well developed and a good place to start.
- Other complex industries must justify their outputs and the quality of their products; health care is no different.
- The dollars spent on collecting data for risk adjustment are trivial compared to the total cost of health care; computerized health information systems can easily generate data for risk adjustment.
- Given the relative dearth of information available about clinicians' performance, we must start with something, even imperfect data.

CABG initiative. In addition to developing clinically detailed risk adjustment algorithms, New York State officials visited CABG program sites that had alarmingly high risk-adjusted mortality rates to see firsthand whether problems existed (Hannan et al. 1990). If providers, payers, and purchasers all recognize the experimental nature of their undertaking, tensions may ease. They could work together to learn how to use the risk-adjusted data productively. The current political climate may not allow time for careful evaluation, but some oversight is essential to determine not only if programs are working as intended but also whether unintended consequences have arisen (e.g., cardiothoracic surgeons are avoiding high-risk patients).

Second, as in all controversies, participants must better understand the goals and concerns of each other. American businesses, for example, often use quality to hone their competitive edge. Clinicians in turn have spent years debating quality measurement, always returning to the difficulty of capturing patients' complexities. For almost a century, emphasis on the idiosyncrasies of individual cases has led clinicians to reject efforts to standardize patient care and even medical record documentation. Physicians typically do not believe that single numerical scores or categories adequately represent complex clinical scenarios. In addition, by focusing on such outcomes as costs, short-term mortality, and length of stay, performance initiatives neglect many outcomes, such as functional status and quality of life, that reflect the goals not only of clinicians but also of patients. With value-based purchasing looming, out of necessity clinicians are learning more about the desire of purchasers and payers to quantify quality to enable more prudent, better-informed decisions. These groups reasonably no longer accept vague promises about quality monitoring without concrete evidence of results. In turn, purchasers and payers should involve clinicians in designing risk-adjusted quality indicators and testing their validity and reliability.

Third, given the uncertainty surrounding the meaning of risk-adjusted outcomes information, it is critical to evaluate what actions these data can reasonably support. Purchasers' need to act aggressively to control costs must be balanced against legitimate questions about the validity of inferences about comparative provider quality from risk-adjusted data. At a minimum, adoption of appropriate and statistically rigorous analytical techniques is essential (see chapters 10 through 12). As did the Massachusetts Division of Health Care Finance and Policy (see Chapter 2), pausing to rigorously evaluate proposed risk adjustment methods before engaging in performance profiling might produce findings that are sobering but nonetheless better inform decision making about appropriate performance metrics (Shahian et al. 2010). Using risk-adjusted performance data to direct punitive actions against providers seems inappropriate without objective evidence that the data are valid. Even if the information appears methodologically sound, the publicity surrounding the release of provider-specific findings may have

untoward consequences (e.g., clinicians might avoid high-risk patients) that would need to be monitored.

Finally, as the IOM Committee on Regional Health Data Networks (1994, 95) observed almost two decades ago:

> The public interest is materially served when society is given as much information on costs, quality, and value for health care dollar expended as can be given accurately and provided with educational materials that aid interpretation of that information. . . . Public disclosure is acceptable *only* when it: (1) involves information and analytic results that come from studies that have been well conducted, (2) is based on data that can be shown to be reliable and valid for the purposes intended, and (3) is accompanied by appropriate educational material.

While this exhortation urges a cool, scholarly approach, nowadays almost nothing in the health care delivery system occurs in a detached, academic manner. Especially through easy access to the Internet, information about providers' performance is disseminated swiftly and widely. Despite the ostensibly laudable goal of assisting consumers by revealing information, a fundamental irony has emerged: Commercial performance profilers offer few details about their proprietary methods for risk-adjusting outcomes and measuring quality; their approaches are "black boxes." Transparency of the risk adjustment methods chosen to support ACA activities will be essential.

Many factors will determine the future implications of generating and disseminating risk-adjusted outcomes information, such as pay-for-performance schemes, the posting of performance profiles on the Internet, and growing financial pressures on providers and the entire health care delivery system—let alone implementation of the ACA, if it moves forward. The only guarantee is that risk-adjusted outcomes information will continue to raise many questions and generate fractious debates for the foreseeable future. However, we are not the first to confront such issues.

Striking historical precedents anticipate the controversies of our times about measuring provider performance and promoting change. I therefore close the book by returning to Florence Nightingale, William Farr, and hospital mortality statistics and the story begun in Chapter 1. (The following text is adapted from Iezzoni 1996.)

Nightingale, Farr, and Hospital Death Rates

In 1863, Florence Nightingale published the third edition of her *Notes on Hospitals*, proposing reforms that she believed would improve the quality of hospitals and patient outcomes. To bolster her arguments about the direction of reforms, she included a table listing death rates in 1861 at 106 hospitals in

England (see Exhibit 1.3). Nightingale (1863, 4) drew attention to the startling mortality rates at 24 London hospitals:

> We have 24 London hospitals, affording a mortality of no less than 90.84 per cent., very nearly every bed yielding a death in the course of the year. . . . Here we have at once a hospital problem demanding solution . . .
>
> Facts such as these (and it is not the first time that they have been placed before the public) have sometimes raised grave doubts as to the advantages to be derived from hospitals at all, and have led many an one to think that in all probability a poor sufferer would have a much better chance of recovery if treated at home.

Subtly embedded in Nightingale's commentary, although obvious from the table, is the method used to calculate these death rates—one that raised questions about interpreting these figures. Taken verbatim from his *24th Annual Report of the Registrar-General*, William Farr (Exhibit 18.2) had calculated the death rates as the total number of deaths at the hospital in 1861 divided by the number of patients at the hospital on April 8, 1861. Thus, the numerator reflected figures from an entire year, whereas the denominator encompassed a single day. Farr had computed death rates per occupied hospital bed—or deaths per person-year in hospital—not death rates per total number of hospitalized patients. Not surprisingly, ostensible hospital death rates fell considerably when calculated as the annual number of deaths divided by the total number of inpatients treated during the year. Using this method, mortality rates for 1861 in the general wards at 14 London hospitals averaged 9.7 percent (Statistical Society 1865).

Cynical modern observers might think that Farr and Nightingale intentionally skewed statistics to bolster political arguments. In the 1860s, however, little consensus existed on statistical techniques, let alone how to calculate hospital death rates. Victorian statisticians emphasized subject matter rather than methods, accepting "men of little mathematical ability" into their field (Eyler 1979, 19). Today's standard statistical techniques and approaches toward error and uncertainty (see chapters 10 through 12) were not introduced until many decades later.

Hospitals calculated death rates in different ways to suit their particular goals (Woodward 1974; Bristowe and Holmes 1864). By changing the specification of the numerator, hospitals could modulate their

EXHIBIT 18.2
William Farr

Source: The Boston Medical Library in the Francis A. Countway Library of Medicine. Used with permission.

apparent death rates. According to the 1863 Privy Council report (Bristowe and Holmes 1864, 527):

> In the majority of hospitals, it is . . . the custom to reckon among their deaths those who have been brought dead to the institution; but there are many hospitals where such cases are not reckoned, and there are some indeed where even those who die within 24 hours are, on the ground that they were moribund at the time of admission, excluded from computation.

Admission practices affected death rate comparisons between urban and provincial hospitals. Many provincial hospitals explicitly refused people with phthisis (consumption), people with fevers, and the "dead or dying," whereas urban facilities accepted everyone. Urban facilities objected to comparisons with outlying hospitals that excluded such cases. As the 1846 Glasgow Royal Infirmary report stated, "The reception of moribund cases greatly swells the number of deaths recorded in the Hospital, and very materially increases the proportionate mortality thereby producing misconceptions in the public mind" (Woodward 1974, 135).

Farr and Nightingale faced criticism primarily because of the denominator they used in their calculations. Nonetheless, in the mid-nineteenth century, some considered the number of deaths per bed a useful indicator of a hospital's productivity—another way of showing charitable donors they were getting their money's worth.

Today's media would swiftly decry hospital mortality rates of more than 90 percent. My review of indexes to *The Times* from 1861 through 1865, however, found few articles about hospitals and none about controversies over Nightingale's book and hospital mortality statistics. Nevertheless, a raucous debate immediately erupted in the London *Medical Times and Gazette* and the *Lancet;* the major critics were men practicing at urban hospitals (Eyler 1979).

An anonymous reviewer (1864a, 129) of *Notes on Hospitals* began, "It is sad to see a work of so much value—full of such useful information—disfigured by a few serious and elementary mistakes. Much as all Medical men must appreciate the philanthropic labour of its authoress, it is a false kindness to pass erroneous views without protest." The reviewer observed that because the mortality rate table came from the Registrar-General, "perhaps Miss Nightingale can hardly be held responsible for it," but he nonetheless excoriated the methodology (Anonymous 1864a, 129):

> The inmates of a single day are balanced with the deaths of a whole year, and no wonder the results are "striking enough." It is to be hoped there are valid reasons for giving to the world what seems to us a simple piece of arithmetical legerdemain. Surely it is the very essence of percentages and of averages (both, we believe, fruitful sources of

> error), that the figures dealt with should stand on one and the same bottom, and that deaths for one year should be compared with admissions or discharges for that period, and no other. There is something audacious in the last column of this table, where twenty-four London Hospitals are accredited with a "mortality per cent. on inmates" of 90.84. No doubt it will be said this is the quotient of the figures employed; but we entirely deny their validity and the accuracy of the impression thus conveyed. The problem as here put is exactly that so often asked of forward schoolboys—What is the quotient of a hundred apples divided by fifteen red herrings?

John Bristowe (1864, 492), a prominent London physician, slyly caricatured Farr's arithmetic choices, showing that hospital recovery rates calculated using Farr's methods would range from 899.5 percent to 953 percent. Timothy Holmes (1864, 365), a London surgeon, indicated that by Farr's method, one hospital had a mortality rate of 130 percent.

Another anonymous critic (1864b, 187) objected to the absence of risk adjustment, viewing comparisons between inner-city and rural hospitals as hopelessly flawed. "Any comparison which ignores the difference between the apple-cheeked farm-laborers who seek relief at Stoke Pogis (probably for rheumatism and sore legs), and the wizzened [sic], red-herring-like mechanics of Soho or Southwark, who come from a London Hospital, is fallacious." Bristowe (1864, 492) concurred:

> Has Dr. Farr . . . really overlooked the differences in relative severity of cases admitted into his different classes of Hospitals, the different relative length of stay of their inmates, the different numbers of patients treated in them in relation to the numbers of constantly-occupied beds? Has he no suspicion that his death-rate is determined almost wholly by these causes?

Bristowe (1864, 491) also questioned how the public would interpret Farr's death rates: "That Dr. Farr understands the mathematical meaning of his figures no one will doubt; but that the majority of his readers understand them neither in this sense nor in any other, and are utterly misled by them, is certain." Bristowe (1864, 492) directly challenged the motivations of Farr and Nightingale, stating that when they "try to mislead others into the belief that the unhealthiness of Hospitals is in proportion to Dr. Farr's death-rates of Hospitals, we are bound to protest against the whole matter as an unfounded and mischievous delusion."

Two weeks after the first review of *Notes on Hospitals*, the *Medical Times and Gazette* published Farr's response. He took exception to an anonymous reviewer "who could treat a lady roughly" (Farr 1864a, 186), although he later accurately acknowledged that Nightingale was "well able to defend herself" (Farr 1864b, 421). Farr argued that if hospitals would provide accurate figures concerning cases treated and deaths, few disputes would arise. In

several back-and-forth rebuttals of his critics over about two months, Farr did not refute specific attacks on his calculation. Instead, he emphasized fundamental reservations about most death rate calculations (Farr 1864a, 186):

> This [Farr's approach] is one method; there is another which is less correct, but more common. The deaths are divided by the mean number of cases admitted and discharged. . . . The defect of this method lies in this: it does not take the element of time into account, which is important, as it so happens that cases are scarcely ever admitted as in-patients of Hospitals at their origin, and that many cases are discharged from Hospital before they have terminated.

Farr (1864a, 186) wanted to hold constant the window of observation, saying, for example, that it was unfair to compare death rates at St. Thomas's in London, which had an average inpatient stay of 39 days, with rates at two Dublin hospitals that had average stays of 27 days. At least, Farr argued, his calculation clearly specified what it aimed to capture.

At the end of the heated letter exchange between Farr and his critics, however, Bristowe (1864, 492) made perhaps the key point: "If Dr. Farr had made his calculations about Hospitals in a tentative spirit, with the object of ascertaining whether they were likely to lead to any useful results, he would have acted in a way to which no exception could have been taken."

One year later, the Statistical Society (1865) with Farr as treasurer published hospital death rates for 1863, calculated as: (annual deaths) ÷ (annual admissions + [patients at the beginning − those at the end of the year]). The publication noted that hospital stays were long, averaging 30 days among 14 London hospitals. Despite this methodological shift, Farr continued using statistics to urge reform, writing to Nightingale in 1864, "What are figures worth if they do no good to men's bodies or souls?" (Diamond and Stone 1981, 70).