Research Evidence on the Validity of Risk-Adjusted Mortality Rate as a Measure of Hospital Quality of Care

J. William Thomas
Timothy P. Hofer
University of Michigan

For more than 10 years, reports comparing quality of care in hospitals have been disseminated to the public. The most commonly used measure in these reports is hospital mortality rate. Despite the resources devoted to analyzing and disseminating mortality data, little attention has been given to the question of their validity as a quality measure. In this article, the authors synthesize findings from 18 articles identified as providing information relevant to this issue. From this review, the authors find evidence that poor quality care increases patients' risk of mortality and that, on average, quality of care provided in hospitals identified as high–mortality rate outliers is poorer than that provided in low–mortality rate outlier hospitals. Nevertheless, a clear conclusion from these studies is that when used as a measure of quality for individual hospitals, risk-adjusted mortality rates are seriously inaccurate. Publication of hospital mortality rates misinforms the public about hospital quality.

In 1993, the Health Care Financing Administration (HCFA) stopped releasing its annual reports on hospital-specific Medicare mortality rates. However, motivated primarily by purchaser concerns about the quality of provider networks, an increasing number of other organizations—state health data agencies, commercial health data vendors, purchaser coalitions, and, since 1991, national news magazines—now periodically publish data that compare hospitals on quality-of-care performance. Risk-adjusted mortality rate is the most

This article, submitted to *Medical Care Research and Review* on November 6, 1997, was revised and accepted for publication on July 20, 1998.

Medical Care Research and Review, Vol. 55 No. 4, (December 1998) 371-404
© 1998 Sage Publications, Inc.

from the SAGE Social Science Collections. All Rights Reserved.

commonly used indicator of quality in these reports. Published mortality rate data are almost always controversial, with many hospitals charging that the numbers are misleading and biased. Defenders of the reports counter that while mortality rates are not perfect measures of quality, no better outcome measures are available at this time. Interestingly, relatively few attempts have been made to ascertain whether or not these measures are valid indicators of hospital quality of care.

Do hospitals that deliver poor-quality care experience higher risk-adjusted mortality rates than other hospitals? Can patients be confident that hospitals with mortality rates significantly lower than expected are actually good-quality providers? For any two hospitals, is difference in risk-adjusted mortality rates indicative of differences in quality performance? In this article, we present a review of 30 years of research and we attempt to answer these questions.

NEW CONTRIBUTION

The health services literature includes scores of articles related to the use of hospital mortality rates as indicators of quality—articles that promote specific types of mortality rate measures, articles that propose risk-adjustment methodologies, articles that describe benefits of reporting hospital mortality rates, and so forth. However, very few of these studies address what is probably the most important issue in this context—whether the hospital mortality rate data *mean* anything; whether or not mortality rates provide *valid* information with which to judge the performance of hospitals. In this article, we synthesize findings from 18 articles that we have identified as providing information relevant to this issue, and we attempt to answer the questions posed in the preceding paragraph.

BACKGROUND

In 1968, Roemer, Moustafa, and Hopkins proposed the "severity-adjusted death rate" (SADR) as an indicator of hospital quality of care. Having observed that hospital mortality rates were highly correlated ($r = 0.79$) with average lengths of stay (ALOS), Roemer, Moustafa, and Hopkins (1968) used ALOS, considered a proxy measure of case mix, to adjust for differences among hospitals in types and severity of patients treated. To test the validity of the SADR as a measure of hospital quality performance, Roemer, Moustafa, and Hopkins (1968) partitioned a sample of hospitals into quartiles on the basis of quality-associated characteristics such as approved residency programs, intensive care units, and blood banks. Consistent with expectations, crude

mortality rates were found to be lowest in least technologically sophisticated quartiles of hospitals and highest in the most sophisticated, since it was presumed that the more complex and severely ill cases were treated in the more technologically advanced hospitals. With the SADR measure, a nearly opposite pattern was observed—the lowest adjusted death rates occurred in the most technologically sophisticated hospitals. This finding was consistent with the assumption that more advanced hospitals provided better quality care, and it was interpreted as evidence that the SADR was a valid measure of hospital quality. However, subsequent research in other settings contradicted this conclusion (Lave and Lave 1971; Goss and Reed 1974), and the SADR was never used outside of the research community.

Throughout the 1970s, researchers and policy makers continued to call for measures of quality capable of focusing on the "bottom line" of medical care, outcomes, rather than on structural factors or processes with which care was delivered (Institute of Medicine [IOM] 1974; Jacobs and Jacobs 1974; Brook et al. 1977). In the 1980s, purchasers too began demanding data on risk-adjusted outcomes to support efforts to "buy right" (McClure 1985). However, it was nearly two decades after publication of Roemer, Moustafa, and Hopkins (1968) before analytic methods necessary for adjusting hospital outcomes would become widely available. HCFA's 1986 release of comparative hospital mortality data demonstrated that such analyses were feasible. The HCFA report raised a host of methodological questions (Blumberg 1987; Vladeck et al. 1988; Jencks, Williams, and Kay 1988; DuBois 1989), such as whether rates should include only deaths that occur in the hospital, whether risk measurement should be patient specific or hospital specific, and whether administrative databases were adequate for risk measurement. Dozens of government, academic, and commercial researchers took up these challenging issues, and by the end of the 1980s, reports comparing hospital performance in terms of mortality and other outcomes were being published annually by state health data agencies, state hospital associations, business coalitions, and commercial health data vendors. By 1991, hospital performance rankings were appearing annually in national magazines (Green et al. 1997; Hill, Winfrey, and Rudolph 1997).

Today, information on hospital quality performance is readily available to purchasers and to the public. However, more than a decade after the first HCFA report, concerns continue to exist about the validity of measures used to assess hospital quality. Such concerns are regularly fueled by responses to new data releases. Hospitals identified as providing superior quality care typically praise the information, stating proudly that the data provide testimony to their institutions' long-standing commitments to continuous improvement in patient care. Hospitals identified as poor-quality providers

typically criticize the data as incomplete and inadequately adjusted for casemix and severity (Berwick and Wald 1990). Do differences in hospitals' risk-adjusted mortality rates indicate real differences in quality of care, or are the rates more reflective of factors outside of the control of hospitals? One would suspect, given the number of hospital quality performance reports that have been published during the past decade, that the accuracy of this type of publicly reported information would have been investigated thoroughly. Such is not the case, however.

For a 1990 article on quality indicators, Sisk et al. located only four publications that addressed the question of validity for mortality rate measures, and three of these focused on HCFA's initial mortality data reports:

- a 1987 New York State Department of Health (NYSDOH) analysis of hospital medical records that identified quality-of-care problems in only 3 percent of cases treated in HCFA high–mortality rate hospitals,
- a 1988 U.S. Congress, General Accounting Office (USCGAO) study that found possible or definite quality problems in only 6 percent of hospitals identified as outliers in HCFA's 1986 report, and
- another New York study (Hannan and Yazici 1988) that reported finding fewer quality problems in records sampled from HCFA high–mortality rate outlier hospitals than in cases sampled from other hospitals.

The fourth article cited by Sisk et al. (1990) was a RAND study (DuBois et al. 1987), in which samples of medical records, selected from high–mortality rate and low–mortality rate outliers in an investor-owned chain of hospitals, were reviewed for quality of care. According to Sisk et al. (1990), this study found that "high mortality hospitals were significantly more likely than low mortality hospitals to have preventable deaths."

On the basis of these articles, Sisk et al. (1990) concluded, "[T]aken together, the results of the few studies that have attempted to validate analyses of hospital mortality suggest caution in using mortality rates as an indicator of quality" (p. 265). In this article, we reconsider the findings of DuBois et al. (1987) cited by Sisk et al. (1990), and we identify two other "early" articles (Williams 1979; Knaus et al. 1986) that contain evidence related to the validity of mortality rate indicators. To these studies, we add more than a dozen others published in, and subsequent to, 1990 that provide evidence related to the validity of risk-adjusted mortality rates as measures of hospital quality. Interestingly, after examining this larger and more recent body of research, our conclusions remain similar to those of Sisk et al. (1990).

In presenting this literature, we organize the studies as follows:

1. Articles that examine validity empirically, focusing on hospital-level relationships between risk-adjusted mortality rates and process-based measures of quality.
2. Articles that examine conditions affecting validity. This group includes
   - articles that provide evidence about population-level relationships between quality of care and patients' mortality risk,
   - articles that focus on potential systematic measurement errors associated with inadequate risk adjustment, and
   - articles that quantify the potential for random measurement error when using mortality rates to evaluate hospital quality performance.
3. Articles that infer validity from observed behavioral responses to public release of hospital mortality rate data.

Table 1 lists articles reviewed in each of these categories.

ARTICLES THAT EXAMINE VALIDITY EMPIRICALLY

A valid measure of provider quality of care should be able to distinguish accurately between providers that are delivering care of acceptable quality and those delivering poor-quality care. Beyond this simple statement, however, the issue of validity for measures of quality is not at all straightforward. A variety of different approaches have been proposed for assessing validity, and these different approaches are usually referred to as different *dimensions* of validity. In a recent chapter, Daley (1994) describes eight dimensions: face validity, content validity, construct validity, convergent validity, discriminant validity, criterion validity, predictive validity, and attributional validity. Each of the empirical validation studies reviewed below uses methodologies based on one of these dimensions.

Criterion validity, construct validity, and *convergent validity* all concern the same questions: How likely is the candidate measure to yield information that is consistent with other accepted measures of the concept? What are appropriate measures of quality of care with which risk-adjusted mortality rates can be compared? As Donabedian (1988) notes,

> All assessments of quality are based . . . on hypotheses concerning the interrelationship among structure, process, and outcome; the assessments are valid only to the extent the hypotheses are verifiable . . . outcomes that are not known to be the consequences of antecedent care cannot be used to assess the quality of that care. (P. 177)

Thus, degree of validity for an outcome-based measure of quality should be proportional to the strength of evidence linking better measured outcomes

(Text continues on p. 380)

TABLE 1 Summary of References on Validity of Risk-Adjusted Mortality Rates as Indicators of Hospital Quality

Type of Validation	*Authors*	*Methodology: Validity of Risk-Adjusted Mortality Rates*	*General Findings: Validity of Mortality Rates*
Empirical-correlational validity	Williams (1979)	Hospital-level regression of quality-related factors on perinatal mortality rate; 504 hospitals	Mortality rates lower in hospitals with good structural and process-of-care characteristics
Empirical-correlational validity	Thomas (1991)	Hospital-level correlations of condition-specific mortality rates with peer-review-judged quality problem rates; 10 diagnoses, 50,000 cases, 42 hospitals	Significant hospital-level relationship between mortality rate and quality for cardiac surgery, acute myocardial infarction (AMI), and pneumonia, but not for seven other conditions
Empirical-correlational validity	Hartz et al. (1993)	Hospital-level correlations of overall mortality rates with peer-review-judged quality of care problems; 2 million cases, 4,132 hospitals in 38 states	Significant hospital-level relationship between mortality rate and quality in 14 states, but not in 28 other states studied
Empirical-predictive validity	Knaus et al. (1986)	Comparison of low-outlier intensive care unit (ICU) to high-outlier ICU on quality factors determined through clinical management audit	Worst ICU had poor staff coverage and poor communications between nurses and physicians
Empirical-predictive validity	DuBois et al. (1987)	Comparison of quality scores for AMI, congestive heart failure (CHF), and pneumonia cases sampled from six high-outlier and six low-outlier hospitals; hospital-level risk-adjustments	No differences in quality measured using explicit criteria; preventable deaths higher in high outliers for pneumonia cases, but not different for CHF and AMI cases
Empirical-predictive validity	Park et al. (1990)	Comparison of mean explicit-criteria quality scores between high-outlier and nonoutlier hospitals, for 1,126 CHF and 1,150 AMI cases; 1,264 hospitals from four states	With demographic risk-adjustment model, no differences in quality scores by outlier status; with better risk models, lower quality score in high outliers for CHF but not for AMI patients

Empirical-predictive validity	Hannan et al. (1990)	Comparison of mean generic-criteria quality scores for 60 coronary artery bypass graft (CABG) deaths sampled from three high-outlier and four low-outlier hospitals	Rate of quality problems 10 times greater in high-outlier hospitals than in low-outlier hospitals
Empirical-predictive validity	California Office of Statewide Health Planning and Development (COSHPD) (1996)	Comparison of explicit-criteria quality scores for 1,005 AMI cases sampled for 10 high-outlier hospitals, 10 low outliers, and 10 nonoutliers	Process compliance significantly better in low-outlier hospitals on some criteria, but not on others
Evidence for valid Variance	Kahn et al., *Measuring Quality* (1990)	Related explicit, condition-specific quality criteria to mortality risk for 14,000 cases for AMI, CHF, pneumonia, stroke, and femur/pelvic fracture cases	Poor quality increases mortality risks by 74% for CHF, 25% for AMI, and 36% each for stroke and pneumonia
Evidence for valid variance	Rubenstein et al. (1990)	Related structured implicit quality scores to mortality risk for 1,197 records sampled from records of Kahn et al., *Measuring Quality* (1990)	Patients receiving poor-quality care are more than twice as likely to die as other patients
Evidence for valid variance	Thomas (1991)	Related Medicare peer review quality judgments to mortality risks for 50,000 cases in 10 clinical conditions in 42 Twin Cities hospitals	Significant relationships between quality and mortality risk for cardiac surgery, AMI, arrhythmia, pneumonia, and femur/pelvic fractures
Evidence on systematic-error variance	Iezzoni et al., *Predicting Who Dies* (1995)	Compared measured mortality risks from four different severity systems for 12,000 AMI patients 20% of patients	Some pairs of severity systems assigned very different scores to more than
Evidence on systematic-error variance	Iezzoni et al., *Using* (1995)	Compared measured mortality risks from 11 severity systems for stroke patients in 27 hospitals	Hospital mortality rate rankings vary for about one third of hospitals depending on severity system used
Evidence on systematic-error variance	Landon et al. (1996)	Compared measured mortality risks from 14 severity systems for 8,000 CABG surgery patients in 38 hospitals	Hospital mortality rate performance rankings were generally consistent among severity methods for 33 hospitals, but not for 5 others

(continued)

TABLE 1 Continued

Type of Validation	*Authors*	*Methodology: Validity of Risk-Adjusted Mortality Rates*	*General Findings: Validity of Mortality Rates*
Evidence on systematic-error variance	Iezzoni et al. (1996)	Compared measured mortality risks from 14 different severity systems for 18,000 pneumonia cases in 105 hospitals	Hospital mortality rate rankings vary for about one third of hospitals depending on severity system used
Evidence on systematic-error variance	Hannan et al. (1992)	Comparison of mortality risks from two different severity systems for 23,000 CABG patients treated in 29 hospitals	Mortality risk predictions developed with detailed clinical findings are significantly more accurate than those from an administrative data set
Evidence on systematic-error variance	Hannan et al. (1997)	Compared mortality risk predictions from clinical and administrative data sets for CABG patients in 31 hospitals	Part of discriminatory power of administrative statistical models is the result of miscoding postoperative complications as coexisting illnesses
Evidence on random-error variance	Park et al. (1990)	Used simulation to investigate random variation in observed mortality when all hospitals have same underlying rate	56% to 82% of difference in mortality rates between high-outlier hospitals and nonoutliers could result purely from random variation
Evidence on random-error variance	Luft and Romano (1993)	Used multiple years of data on CABG mortality rates in California to investigate year-to-year variation in hospital mortality rates	Mortality rates in high-outlier hospitals averaged 31% higher than expected levels, 2 years after high-outlier designation; no hospital identified as high outlier in 1987 was still a high outlier in 1989

Evidence on random-error variance	Hofer and Hayward (1996)	Simulated state hospital system to investigate specificity and predictive error when identifying poor-quality hospitals using mortality rates of medical patients	Sensitivity of mortality measure for identifying poor-quality hospitals was less than 10%; predictive error ranged from 76% to 84%
Evidence on random-error variance	Zalkind (1996)	Simulation study testing multiple scenarios for accuracy of identifying poor-quality hospitals using mortality rates	With volume of 200 patients per hospital, sensitivity averages 10% to 20% for identifying poor-quality hospitals; predictive error averages 60%
Evidence on random-error variance	Thomas and Hofer (forthcoming)	Uses six-parameter analytic model to develop exact measure of sensitivity and predictive error for mortality rate measure of quality	Using estimates from Kahn et al., *Measuring Quality* (1990) as model parameters, found sensitivity < 12% and predictive error > 60%
Inferring validity from behavioral responses	Rosenthal, Quinn, and Harper (1997)	Documents decline in average mortality rates in eight diagnoses among 30 hospitals in northeastern Ohio following publication of mortality rates	During 2-year period, average mortality decreased from 7.5% to 6.5%; rate of decline was statistically significant for CHF and pneumonia
Inferring validity from behavioral responses	Hannan et al., *Improving the Outcomes* (1994)	Documents decline in CABG mortality rates among New York hospitals following publication of mortality rate data; declines occurred among high outliers, low outliers, and nonoutliers	Observed a 41% decline in CABG mortality rates following publication of hospital mortality rates

with superior processes and worse measured outcomes with inferior processes. With criterion validity, the standard with which the measure is compared is one that is considered unequivocally valid, a "gold standard." Construct validity (or correlational validity) is the same as criterion validity, except that the comparison is made with one or more measures thought, but not known, to be highly valid. This is the validation criterion employed in three of the empirical validation studies cited below (Williams 1979; Thomas 1991; Hartz et al. 1993), each of which examines relationships between hospitals' risk-adjusted mortality rates and a process-based measure of quality performance.

To have *predictive validity*, a measure should perform well in predicting the occurrence of an attribute (e.g., good-quality care) or, equivalently, in discriminating among occurrences of the attribute (*discriminant validity*). With only the three exceptions cited above, all of the empirical validation studies cited below focus on predictive validity (Knaus et al. 1986; DuBois et al. 1987; Park et al. 1990; Hannan et al. 1990; California Office of Statewide Health Planning and Development [COSHPD] 1996). It is important to note that in the context considered here, the criterion refers to the ability of a hospital mortality rate measure to predict *hospital quality performance*. Evidence that a risk model accurately predicts mortality does not, as has sometimes been suggested (e.g., DesHarnais et al. 1988), establish predictive validity for risk-adjusted mortality rates.

In this section, we describe findings from studies that have investigated relationships between risk-adjusted mortality rates and measures of hospital quality performance that are in some way based on process-of-care judgments. The studies are presented chronologically by validity measurement method.

CORRELATIONAL VALIDITY STUDIES

Predating other risk-adjustment studies by nearly a decade, Williams (1979) employed basic epidemiological methods to develop indirectly standardized perinatal mortality rates for hospital maternity services. By linking birth certificate records of 3.44 million babies delivered in California hospitals in 1960 and 1965-1973 with death certificate records for the 39,000 fetal and neonatal deaths during this period, Williams (1979) developed for each hospital an observed-to-expected perinatal mortality ratio (O/E ratio), where expected rates were based on infant sex, ethnicity, and birth weight. To determine how risk-adjusted perinatal mortality related to quality of care, Williams (1979) performed a stepwise regression analysis of hypothesized quality-related hospital structural and process characteristics on hospital O/E ratios

for the 504 study hospitals. Regression coefficients suggested that, consistent with expectations, risk-adjusted perinatal mortality was lower in hospitals with higher specialist-to-generalist physician ratios and hospitals that routinely measured and recorded infants' Apgar scores. Rates were found to relate to annual volume of deliveries in a U-shaped relationship, declining with increases in volume up to 2,850 annual births, and then increasing with greater volumes. Because this model was estimated by using stepwise techniques and was not tested on independent samples of data, these relationships could reflect some degree of overfitting.

Thomas (1991) examined relationships between hospital mortality rates and quality of care by correlating providers' condition-specific O/E ratios with their condition-specific quality problem rates (percentage of reviewed cases at each hospital identified by Peer Review Organization [PRO] review as involving one or more quality problems).[1] Earlier population-level analyses had indicated valid relationships between mortality risks and quality of care for 5 of 10 clinical conditions studied: pneumonia, cardiac surgery, acute myocardial infarction (AMI), cardiac arrhythmia, and femur/pelvic fractures. For these conditions, it was hypothesized that hospitals with higher rates of identified quality-of-care problems would have higher risk-adjusted mortality rates, and that facilities having few identified quality problems would similarly have low-mortality O/E ratios. Significant ($p < 0.05$) positive correlations, suggestive of valid hospital-level relationships between process and outcome measures, were found for cardiac surgery ($r = 0.55$), AMI ($r = 0.57$), and pneumonia ($r = 0.31$), but not for the other two conditions studied.

A similar analytic approach was used by Hartz et al. (1993), who examined relationships between PRO-determined quality problem rates and hospitals' risk-adjusted mortality rates for Medicare admissions in 38 states. For 14 of the states, Hartz et al. (1993) found significant ($p < 0.05$) positive within-state correlations, with $r = 0.19$, between hospital quality problem rates and risk-adjusted mortality rates. Among the 6 states having the highest number of PRO reviews, Hartz et al. (1993) found that within-state hospital-level correlations were even higher when more homogeneous groupings of hospitals were analyzed—for example, state/city/county hospitals (mean $r = 0.42$), hospital members of Council of Teaching Hospitals (mean $r = 0.25$). However, for the other 24 states studied, HCFA's 1987 overall risk-adjusted hospital mortality rates did not relate as hypothesized to PRO quality problem rates—correlations were either nonsignificant or negative. One reason for the partially negative findings of both Hartz et al. (1993) and Thomas (1991) could be the reliability of PRO peer review judgments, shown in other studies to be problematic (Goldman 1992; Rubin et al. 1992). Hartz et al. (1993) reported that among the 38 state PROs submitting data for their study, physician-

confirmed quality problem rates, across all diagnoses and hospitals, ranged from 0.0003 (New Jersey) to 0.3846 (Puerto Rico) and averaged 0.0373.

PREDICTIVE VALIDITY STUDIES

In this section, we review five studies in which sets of superior-quality hospitals and poor-quality hospitals were identified on the basis of risk-adjusted mortality rates, and then these quality predictions were evaluated by comparing processes of care between the two types of outliers. The first listed study, Knaus et al. (1986), developed information on hospital quality through clinical management audits of hospitals participating in the study. Each of the other four studies based comparative quality judgments on the results of medical record reviews from samples of high-outlier hospitals and samples of nonoutlier or low-outlier hospitals.

After ranking 13 hospital intensive care units (ICUs) on the basis of risk-adjusted (using APACHE II) mortality rates, Knaus et al. (1986) performed clinical management audits of each of the subject units. The hospital having the lowest O/E ratio (0.59) was found to have several characteristics presumed to be associated with good quality—for example, a full-time ICU director and 24-hour in-unit physician coverage. The ICU with the worst risk-adjusted mortality rate (O/E = 1.58) did not have 24-hour in-unit physician coverage and did not have a full-time director. Most important, the unit was found to suffer from very poor communications between physicians and nursing staff. Although the Knaus et al. (1986) study might be considered to support the validity of risk-adjusted mortality rate as a measure of quality, it represents weak evidence, both because of the small number of hospitals surveyed and because of the post hoc nature of the clinical management audits.

Of the four studies reviewed by Sisk et al. (1990), the only one supportive of the validity of mortality rate as a quality indicator was the article by DuBois et al. (1987). With data from a chain of 93 proprietary hospitals, DuBois et al. (1987) developed a hospital-level model to predict the number of in-hospital deaths as a function of four casemix indicators: percentage of patients older than 70, percentage of admissions from the emergency department, percentage of admissions from nursing homes, and Medicare hospital casemix index. Comparing the hospitals' observed mortality rates with rates predicted by the model, DuBois et al. (1987) identified 9 low outliers (presumed good quality) and 11 high outliers (presumed poor quality). Samples of medical records for pneumonia, stroke, and AMI patients were obtained from six hospitals in each group. The records were audited for quality using explicit condition-specific criteria and also using physicians' implicit judgments about preventability of death. Based on compliance with explicit condition-specific criteria, results

showed no differences between high- and low-outlier hospitals for any of the conditions studied. When quality of care was judged implicitly, no differences in scores were observed between high- and low-outlier hospitals for AMI patients or stroke patients. However, for pneumonia patients, preventable death rates were judged to be higher in the group of high-outlier hospitals than in the group of low-outlier hospitals. Despite negative explicit-review results for all cases and negative implicit-review results for AMI and stroke cases, DuBois et al. (1987) interpreted this pneumonia finding as evidence supportive of the validity of risk-adjusted mortality rates as quality indicators. The study and this conclusion have been criticized on several grounds, one of which was the low level of interrater reliability for physicians performing implicit reviews. Although reliability statistics were not reported in the original article, in a later article, DuBois (1989) indicated that kappas were as low as 0.11 for pneumonia cases reviewed.

The DuBois et al. (1987) risk model was similar to the 1986 HCFA mortality model in that it was designed to predict number of deaths as a function of hospital volume and casemix statistics; that is, hospital was the unit of analysis. Although not recognized at the time of the DuBois et al. (1987) study, researchers subsequently learned that hospital-level risk-adjustment models produce severely biased predictions, and that mortality risk adjustments instead must be based on patient-level models (Hadorn et al. 1993). With patient-level models, the expected number of deaths for a hospital is calculated by summing individual patient mortality probabilities across patients treated at a hospital. This is the approach used in the 1990 study by Park et al., who, like DuBois et al. (1987), wished to determine whether hospitals targeted as high–mortality rate outliers were actually delivering poorer quality care than other hospitals.

The initial risk model used by Park et al. (1990) considered only demographic (age, sex, race) predictors of mortality. Comparing high-outlier and nonoutlier hospitals identified with this model, Park et al. (1990) found no significant differences in quality—measured as degree of compliance with process criteria—for either congestive heart failure (CHF) or AMI cases. They noted, however, that for both AMI and CHF patients, mean quality scores in high-outlier hospitals were as high or higher than those in nonoutlier hospitals. They also noted that more than 80 percent of differences in adjusted mortality rates between high-outlier and low-outlier hospitals were not attributable to quality but to residual severity differences and random (binomial) variation in the outcome. For a subsequent set of analyses, Park et al. (1990) compared quality of care between high-outlier and low-outlier hospitals identified in HCFA's 1988 mortality report (1986 data), and between high-outlier and low-outlier hospitals identified by using 3 years rather than 1 year of Medicare data. With each of these new approaches for identifying mortality

rate outlier hospitals, Park et al. (1990) found that the average quality-of-care score for CHF cases treated in high-outlier hospitals was poorer than the average in low-outlier hospitals. For AMI patients, quality score differences between groups of high-outlier and low-outlier hospitals were not significant, but patterns were considered suggestive of better quality in low-outlier hospitals.

Among all of the organizations currently distributing public reports on hospital quality-of-care performance, we are aware of only two that have made efforts to investigate the validity of the measures presented. The first of these is NYSDOH, which in 1989 began releasing annual reports on the mortality rate performance for New York hospitals performing coronary artery bypass graft (CABG) surgery (Hannan et al. 1990). Since 1992, the department has also produced surgeon-specific reports (Hannan et al., *Improving the Outcomes*, 1994). In the original article of an ongoing sequence, Hannan and colleagues (1990) described their use of data from a cardiac surgery registry to model patients' CABG mortality risks in terms of cardiac system status, complications and comorbidities, and demographics. For the 28 hospital programs evaluated, crude CABG mortality rates ranged from 2.2 percent to 14.3 percent. Comparing these figures to the expected numbers of deaths, 3 of 28 programs were found to have significantly fewer deaths than expected, and 4 were found to have significantly higher mortality rates than expected. Through a contract with the NYSDOH, the Island PRO examined the medical records of 40 patients who had died following surgery at the high–mortality rate hospitals and 23 who had died at low–mortality rate hospitals. Applying generic process-of-care criteria, the reviewers found 45 percent of the deaths in high-outlier hospitals to have associated quality-of-care problems, while only 4.35 percent (one case) experienced questionable quality care in the low-mortality hospitals. Hannan et al. (1990) do not indicate whether medical record reviewers were aware of hospital outlier status when evaluating care documented in the charts.

In conjunction with California's Hospital Outcomes Project, the state's Office of Statewide Health Planning and Development produces annual statistics on risk-adjusted AMI mortality rates by region and by individual hospital (COSHPD 1996). To investigate the validity of this measure as an indicator of hospital quality, researchers stratified 228 hospitals on the basis of risk-adjusted AMI mortality rates into three categories—better-than-expected mortality (lowest 5 percent), expected mortality, and worse-than-expected mortality (highest 5 percent)—and selected 10 hospitals from each stratum. They then sampled about 325 AMI admissions from each stratum of hospitals. Medical records for the sampled patients were reviewed to assess degree of compliance with hypothesized good processes for treatment of acute myocardial infarctions. Criteria included use of aspirin, thrombolytics, beta blockers,

and heparin; as well as rates of performance of coronary angiography, revascularization procedures, and pulmonary artery catheterization. No significant ($p < 0.10$) differences among hospital mortality rate categories were observed in percentages of patients receiving aspirin therapy or thrombolytic therapy. Although percentages of patients receiving aspirin therapy did not differ among strata, patients in low-mortality hospitals were found more likely than those in other hospitals to receive aspirin therapy within 6 hours of presentation. Patients in low–mortality rate hospitals also were more likely than other patients to receive heparin and to undergo revascularization, coronary angiography, and pulmonary artery catheterization, while patients in high–mortality rate hospitals were more likely to receive beta blockers. The findings were considered to support the validity of risk-adjusted AMI mortality rate as a measure of hospital quality performance (COSHPD 1996).

ARTICLES THAT EXAMINE CONDITIONS AFFECTING VALIDITY[2]

McAuliffe (1984) notes that for any measure m of a concept, its variance σ_m is composed theoretically of three components:

$$\sigma_m = \sigma_v + \sigma_{se} + \sigma_{re},$$

where σ_{re}, represents *valid variance*, the portion of total variance that is related to the concept being studied (e.g., quality differences among hospitals). The remaining variance in the measure consists of two error components: systematic-error variance, σ_{se}, and random-error variance, σ_{re}. The systematic-error variance reflects the influence on the measure of factors that are unrelated to the concept being studied. Random-error variance represents the portion of total variance that remains after accounting for the stable effects of valid variance and systematic-error variance; it is the degree of unreliability in the measure. With this as a framework, we can state conditions for the relative degree of validity of a measure:

- Validity is greater when the ratio σ_v/σ_m is greater.
- Validity is greater when residual σ_{se} is smaller, that is, when a greater proportion of systematic-error variance can be removed, through identification of sources of systematic error, measurement of the influence of those sources on m, and adjustment of m to remove bias from these effects.
- Validity is greater when the proportion of total variance represented by σ_{re} is small, since measures with low reliability *must* have low validity (McAuliffe 1984).

In this review, articles identified as addressing these validity conditions are presented in three groups. Articles in the first group provide evidence related to whether or not $\sigma_v > 0$. The second group of articles concerns the potential for systematic measurement error in hospital mortality rates, that is, whether currently available methodologies adequately control for the effects of casemix and severity differences among institutions. The remaining group of articles focuses on the magnitude of σ_{re} relative to other variance components.

STUDIES OF VALID VARIANCE: EVIDENCE OF POPULATION-LEVEL RELATIONSHIPS BETWEEN QUALITY OF CARE AND PATIENTS' MORTALITY RISK

The clinical literature includes thousands of articles that describe randomized clinical trials or other carefully controlled experiments that document associations between specific clinical interventions and improved patient outcomes. We shall not include such studies in this review, since they focus on very specific interventions in highly selected populations of patients. Instead, we shall focus on articles in which patients' mortality risks were compared by using process measures that are more broadly applicable than those evaluated in randomized clinical trial studies. As Jencks (1995) has noted, this type of validation can be difficult because these types of process measures usually depend on clinician judgments of overall quality, and such judgments can be subjective and unreliable.

Perhaps the most carefully designed investigation available in the literature of relationships between patients' mortality risks and quality of medical care processes was performed in conjunction with RAND's assessment of Medicare's diagnostic related group (DRG)-based Prospective Payment System (Kahn et al., *Comparing Outcomes*, 1990). For each of five conditions, risk models for 30-day mortality were developed by using physiologic and other variables abstracted from patients' medical records (Keeler et al. 1990). Quality reviews were performed on 14,000 cases using explicit, disease-specific process-of-care criteria (Kahn et al., *Measuring Quality*, 1990), and on a subsample of these cases using a structured implicit review methodology (Rubenstein et al. 1990). On the basis of explicit quality review, congestive heart failure (CHF) patients judged as having received poor-quality care were found to be 1.74 times more likely to die within 30 days than patients whose care was acceptable (Kahn et al., *Measuring Quality*, 1990). The research showed that AMI patients with poor care were 25 percent more likely to die than other AMI patients, and that the mortality odds ratios for pneumonia and stroke cases with poor-quality care were both 1.36. Only for patients with hip fractures were Kahn et al. (1990), in *Measuring Quality*, unable to show a sig-

nificant relationship between quality of care and mortality risk (odds ratio = 0.90).[3] When quality-of-care assessments were based on implicit review rather than on explicit criteria, the evidence of a significant relationship between mortality risk and quality of care was even stronger (Rubenstein et al. 1990). For a sample of 1,197 records (across the five conditions studied) that were evaluated using RAND's structured implicit review methodology, cases judged as having received poor-quality care were found 2.08 times more likely to die within 30 days of hospitalization than cases whose care was considered acceptable (Rubenstein et al. 1990). Although not the major purpose of the RAND project, the study provided strong empirical evidence that poor-quality care increases patients' risks of death.

However, findings by Thomas (1991), who also examined population-level relationships between mortality risks and quality of care, were somewhat more equivocal. Thomas (1991) first developed condition-specific models for predicting patients' in-hospital mortality risks using an administrative data set that included all hospital admissions in the Twin Cities, Minnesota, during 1987 and 1988. These models were then applied to estimate mortality risks for nearly 50,000 Medicare admissions that had been reviewed for quality of care by the Foundation for HealthCare Evaluation, Minnesota's Medicare PRO. With these data, Thomas (1991) compared the O/E ratio for patients whose care had been judged of poor quality with the O/E ratio for all patients whose care was judged acceptable. The comparisons showed very strong relationships between patient-level mortality risks and quality of care for patients admitted for cardiac surgery, for AMI, or for cardiac arrhythmia. Somewhat weaker but still significant relationships were noted between quality of care and mortality risks for pneumonia patients and femur/pelvic fracture patients. However, O/E ratios were found not to differ as a function of measured quality for stroke, heart failure, bowel procedure, prostatic disorder, and septicemia cases. Procedures for assessing quality of care used by the Minnesota PRO were quite different from those used by RAND researchers in the study described above (Kahn et al., *Measuring Quality*, 1990; Rubenstein et al. 1990). Because PRO quality determinations could lead to legal and financial sanctions of providers, HCFA designed the PRO quality review process intentionally to minimize the occurrence of false positives. However, despite differences in review methodology, the percentage of patients identified by Minnesota's PRO as having received poor-quality care was the same as that (12 percent) reported by Rubenstein et al. (1990). Furthermore, findings with respect to patient-level relationships between mortality risk and quality reported by Thomas (1991) were consistent with those reported by Kahn et al. (1990), in *Measuring Quality*, for the same conditions.

Why were population-level relationships between quality of care and mortality not significant for *all* conditions examined in these two studies? One plausible explanation is that risk models, especially the administrative data models used by Thomas (1991), did not adequately adjust for differences in mortality risks between patients who had received good-quality care and those whose care was poor. Another possible reason is that mortality may be less sensitive to differences in quality of care for some conditions than for others. While mortality risks of CABG surgery patients may vary greatly with quality of care, for patients with hip fractures, ability to ambulate without assistance is likely to be a more relevant indicator of quality.

STUDIES OF SYSTEMATIC ERRORS IN MORTALITY RATE INDICATORS

The health services literature now includes a large volume of studies that describe methodologies for adjusting mortality rates to account for variation in patient risks (Hadorn et al. 1993; Iezzoni 1994). Nevertheless, concerns still exist about whether even our currently advanced risk-adjustment tools are adequate to support valid comparisons among hospitals. The manner in which such methodologies can introduce systematic measurement error into hospital mortality risk adjustments was illustrated well in Blumberg's (1991) critique of the MedisGroups severity measurement system. Blumberg (1991) compared actual mortality rates with expected mortality rates, with predictions based on MedisGroups' admission severity scores for a group of Medicare beneficiaries hospitalized for AMI. He found that the O/E ratio was 0.77 for patients who were younger than 70 years and 1.41 for patients who were 85 years or older. A similar bias was noted for specific AMI diagnoses: the O/E ratio for patients with principal International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) diagnosis of 410.9 (AMI not otherwise specified) was 1.4, while for 410.7 (subendocardial infarction) it was 0.61. MedisGroups did not explicitly control for either of these factors. As a consequence, risk-adjusted mortality rates calculated on the basis of MedisGroups' admission severity would be inappropriately low for hospitals providing care to young AMI populations and/or populations including disproportionately large percentages of patients with subendocardial infarction. Likewise, hospitals treating older AMI populations and higher percentages of 410.9 patients would have inappropriately high O/E ratios.

Since Blumberg's 1991 study, the MedisGroups system has been refined (Steen et al. 1993). Refinements to improve accuracy of mortality predictions have also been made to other previously available systems (Knaus et al. 1991; Young, Kohler, and Kowalski 1994), and new severity measurement method-

ologies have been introduced (e.g., Romano et al. 1995; Rosenthal and Harper 1994). However, in a set of recent articles, Iezzoni and colleagues present evidence suggesting that the types of problems described by Blumberg (1991) might still be present in risk-adjustment methodologies today (although to a considerably lesser degree). Evaluating four common severity methodologies for predicting AMI patient mortality, Iezzoni et al. (1995), in *Predicting Who Dies*, found that for more than 20 percent of patients studied, mortality risk rankings differed significantly from one severity measure to another. They concluded that "predicting who dies depends on how severity is measured." Iezzoni and colleagues also looked at effects of systematic measurement error on the reliability of hospital mortality rate performance rankings. In analyzing stroke mortality, they applied 11 different severity measures to data from 27 hospitals (Iezzoni et al., *Using*, 1995). For CABG surgery mortality, they compared 14 severity measures with data from 38 hospitals (Landon et al. 1996); and for pneumonia mortality, they examined 14 severity measures using data from 105 hospitals (Iezzoni et al. 1996). Findings in the three studies were generally consistent. As reported in Iezzoni et al. (1996), 73 of the 105 hospitals were identified by all 14 measures as pneumonia mortality rate "inliers" (observed rates not significantly different from expected rates), and 2 other hospitals were uniformly identified as high outliers. However, 14 hospitals were identified as outliers by some of the severity measures, but not by others. In an article summarizing conclusions from these studies, Iezzoni (1997) comments, "Severity measures frequently disagreed about which hospitals had particularly low or high z scores. Agreement in identifying low- and high-mortality hospitals between severity-adjusted and unadjusted death rates was often better than agreement between severity measures" (p. 1600).

Other comparative studies have raised similar concerns. In conjunction with the New York State Health Department's release of data on CABG mortality rates, Hannan et al. (1992) compared relative hospital rankings obtained from analyses using an administrative database and analyses using the state's Cardiac Surgery Reporting System (CSRS). The CSRS is a coronary surgery data registry that includes, in addition to administrative data for each case, information on ejection fraction, whether the surgery represents a reoperation, and whether there is more than 90 percent narrowing of the left main trunk. Hospital mortality rates developed with the two different systems correlated at only 0.75 to 0.80. On the basis of these analyses, Hannan et al. (1992) concluded the following: "[T]o inform consumers of relative quality of hospital care, the differences in hospital ratings between the two systems as well as the potential damage to a hospital's reputation are probably too great to risk using an administrative data base" (p. 903).

In a more recently reported study, Hannan et al. (1997) found that part of the accuracy of CABG mortality predictions developed by using a Medicare administrative database resulted from miscoding of postoperative complications as coexisting illnesses by hospitals. When the complications were removed as mortality predictors in the model, O/E ratios and relative performance rankings changed significantly for the 31 New York hospitals that perform bypass surgery. Hartz and Kuhn (1994) undertook a similar project to Hannan et al. (1992), using a "clinically rich" database and a separate administrative database to develop sets of models for predicting mortality, major complication, and "any complication" for CABG surgery patients. With these models, they risk adjusted CABG patient outcomes in 10 hospitals and compared hospital performance rankings associated with the two different sets of models. None of the rank correlations were significant: 0.48 for mortality, 0.21 for major complications, and –0.14 for "any complication."

STUDIES OF RANDOM MEASUREMENT ERROR OF MORTALITY RATE INDICATORS

The earliest evidence that random variation in hospital mortality rates might make quality inferences problematic was provided by Park et al. (1990). As noted above, Park et al. determined that more than 80 percent of differences in adjusted mortality rates between high-outlier and low-outlier hospitals were attributable to measurement error—to residual severity differences and to random binomial variation in the outcome. As a part of the study, Park et al. (1990) used simulation analysis to determine the proportion of overall variation in hospital mortality rates potentially attributable simply to randomness. Under the null hypothesis that all hospitals had the same underlying age/sex/race mortality rates, they found that, depending on condition (CHF or AMI) and mortality outcome (in-hospital or 30-day), random binomial variation accounted for between 56 percent and 82 percent of differences in observed hospital rates. Analysis of quality-of-care reviews for samples of medical records at high-outlier and nonoutlier hospitals showed that "hospitals targeted with unexpectedly high age-sex-race-disease-specific death rates do not provide lower quality of care than do untargeted hospitals" (Park et al. 1990).

A slightly different conclusion was reached by Luft and Romano (1993), who investigated year-to-year variation in risk-adjusted mortality rates among California CABG surgery programs. With discharge data for the period 1983 to 1989, Luft and Romano (1993) identified high-mortality outlier hospitals based on analysis of consecutive 2-year periods and then investigated whether these same hospitals remained as outliers 2 years later. Risk-

adjusted mortality rates for 1989 were examined for hospitals that had been identified as outliers based on analysis of 1986-1987 data, and similar analyses were done for 1986, 1987, and 1988, focusing on hospital outliers identified through analysis of data from 1983-1984, 1984-1985, and 1985-1986, respectively. Analyses were performed separately using data for high-risk patients (those in the highest quartile of expected mortality) and data for low-risk patients (those in the lowest quartile of predicted mortality). With high-risk CABG patients, mortality rates for hospitals that had been identified as high-outlier hospitals 2 years earlier were found to average 31 percent above expected levels. Mortality rates 2 years later in hospitals identified as low outliers were 28 percent lower than expected. While these findings suggest persistency over time in average risk-adjusted mortality rates for both low-outlier and high-outlier hospitals, the data revealed significant variability among facilities. None of the high outliers selected using 1987 data were still significant outliers in 1989, and casemix-adjusted mortality rates for some outlier hospitals were lower than expected. For low-risk patients, Luft and Romano (1993) found that mortality outlier status was not predictive of future hospital performance.

All of the empirical predictive validity studies reviewed in the section above—DuBois et al. (1987); Park et al. (1990); Hannan et al. (1994), in *Improving the Outcomes*; and COSHPD (1996)—were designed to determine whether, as a group, patients treated in high–mortality rate hospitals were at greater risk of receiving poor-quality care than patients treated in low–mortality rate hospitals. None of these studies attempted to ascertain whether or not individual high-outlier hospitals were delivering poor-quality care. However, this was the specific research question addressed in the three studies that follow, two of which used Monte Carlo simulation and one an analytic model to investigate the accuracy of identification of poor-quality hospitals using risk-adjusted mortality rates. These methodologies allowed analysis of situations in which no casemix differences existed among hospitals—that is, in which risk adjustment was perfect—so that the only factors affecting hospitals' observed mortality rates were quality of care and random binomial variation.

In the first of these studies, Hofer and Hayward (1996) simulated a hypothetical hospital system modeled on hospitals in the state of Michigan—the same number of hospitals (191) and the same distribution of patients per hospital. In this simulated system, patients died at an average rate of 13 percent across all hospitals. Of the hospitals, 10 percent were designated arbitrarily by Hofer and Hayward (1996) as poor-quality providers having high rates of preventable deaths. At the poor-quality hospitals, 12.5 percent of all medical patient deaths were considered preventable. For the other 90 percent of hospitals—the average-quality providers—only 2.5 percent of deaths were

considered preventable. Because of quality differences, medical patients' average mortality risks in poor-quality hospitals was 14.17 percent, compared with 12.87 percent in average-quality hospitals. For each iteration of the simulation model, an observed number of deaths were generated randomly for each of the hospitals, and the 5 percent of hospitals with the highest mortality rates, whether known to be poor-quality or average-quality hospitals, were designated as high outliers. On the basis of more than 100 simulation iterations, Hofer and Hayward (1996) found that only 35 percent of the poor-quality hospitals were correctly identified as high outliers. They also found that 48 percent of the hospitals identified as high outliers were not poor-quality providers: that is, the sensitivity of mortality rate as a hospital quality measure was determined to be 35 percent, and its predictive error was 48 percent. When high-outlier status was defined on the basis of deaths in specific diagnoses (AMI, CHF, stroke, pneumonia) rather than on all medical admissions, because of lower patient volumes per hospital, sensitivity fell to less than 10 percent and predictive error increased to more than 75 percent.

Zalkind (1997) used different assumptions from Hofer and Hayward (1996) in his study, but his conclusions were quite similar. Zalkind (1997) simulated a hypothetical hospital system in which underlying patient mortality risk was 12.0 percent in 5 percent of the hospitals (very poor-quality facilities), 11.5 percent in 20 percent of the hospitals (poor-quality facilities), and an average of 9.8 percent in other hospitals. For simulation runs in which 200 patients were treated in each hospital, he determined that average sensitivity was 9.3 percent for very poor-quality hospitals and 11.0 percent for poor-quality hospitals (trim point at the 95th percentile), and average predictive error was 60 percent. Results were better with runs in which hospitals treated larger patient volumes, but even with 1,000 patients per hospital average sensitivity was less than 25 percent and predictive error was greater than 50 percent.

In the third study, Thomas and Hofer (forthcoming) used a six-parameter analytic model to explore the effects of random binomial variation on hospitals' observed mortality rates. From RAND's Prospective Payment Evaluation study (Rubenstein et al. 1990), they obtained values for three of the model parameters: the proportion of all patients treated who suffer poor-quality care and patients' mortality risks associated with both good-quality and poor-quality care. Thomas and Hofer (forthcoming) referenced published hospital report cards to obtain values for the fourth parameter, hospital patient volume. They obtained estimates for the other two parameters, proportion of hospitals that deliver poor-quality care and patients' relative risk of poor care in those hospitals, from databases of the Texas Foundation for Medical Care, Medicare's PRO contractor for the state of Texas. The analytic model provided an exact measure of the accuracy of mortality rate identification of poor-

quality providers. It showed that fewer than 12 percent of poor-quality hospitals (those in which patients' risks of receiving poor-quality care were 4 times greater than in average-quality hospitals) would be identified as high–mortality rate outliers, and that more than 60 percent of the hospitals identified as high outliers would actually be good-quality providers. Estimates for sensitivity and predictive error for mortality rate identification of poor-quality hospitals were shown to vary with assumptions about hospital volume, impact of quality on patients' mortality risks, and other factors. Thomas and Hofer (forthcoming) concluded, as did Hofer and Hayward (1996) and Zalkind (1997), that under virtually all realistic assumptions, even perfectly risk-adjusted mortality rates are highly inaccurate indicators of hospital quality performance. Thomas and Hofer (forthcoming) did identify one possible exception to this general conclusion. The predictive error in New York State's CABG surgery mortality rate report could be as low as 20 percent—compared with error rates of greater than 60 percent in other mortality data reports—because of both the strong relationship between quality of care and patient mortality risk for CABG surgery patients and the relatively high median volume of CABG surgery in New York hospitals ($N = 530$).

PAPERS THAT INFER VALIDITY FROM OBSERVED BEHAVIORAL RESPONSES TO PUBLIC RELEASE OF HOSPITAL MORTALITY RATE DATA

Neither of the final articles included in this review attempts to demonstrate the validity of mortality rate measures of quality directly. Instead, they present evidence that publication of mortality rate data leads hospitals to improve quality of care. Since such responses would be expected only if the published data were valid, findings from these studies often are cited as evidence from which the validity of mortality rate indicators of quality can be inferred.

Rosenthal, Quinn, and Harper (1997) investigated changes in mortality experience of northeastern Ohio hospitals following public release of performance data by the Cleveland Health Quality Choice program. With data on more than 100,000 consecutive eligible discharges from 30 Cleveland area hospitals, Rosenthal, Quinn, and Harper (1997) noted that average mortality rates across eight diagnoses studied declined during four sequential reporting periods from 7.5 percent to 6.5 percent. For two individual diagnoses, the mortality rate declines were statistically significant—0.50 percent per period for CHF and 0.38 percent per period for pneumonia. The authors commented that "although changes in hospital care were not directly examined, the results suggest that initiatives to examine provider performance may have a beneficial impact on quality of care" (Rosenthal, Quinn, and Harper 1997).

The approach of Rosenthal, Quinn, and Harper (1997) was similar in concept to an earlier evaluation of the New York State CSRS performed by Hannan and colleagues (Hannan et al., *New York's Cardiac Surgery*, 1994; Hannan et al., *Improving the Outcomes*, 1994). Four years after NYSDOH began releasing hospital-specific CABG surgery survival rate data to the public, actual mortality associated with the procedure had dropped 21 percent, from 3.52 percent statewide in 1989 to 2.78 percent in 1992 (Hannan et al., *Improving the Outcomes*, 1994). During this same period, measured severity of cases treated increased 25 percent, so that, on a risk-adjusted basis, risk-adjusted CABG mortality in New York dropped from 4.17 percent to 2.54 percent—a 41 percent decline. Of the five hospitals identified as high-mortality outliers in 1989, all demonstrated lower rates in 1990, 1991, and 1992. Hannan et al. (1994), in *New York's Cardiac Surgery*, trichotomized the 30 New York hospitals that performed CABG surgery in 1989 into a higher-than-expected risk-adjusted mortality group, an as-expected group, and a lower-than-expected group. From 1989 to 1992, risk-adjusted CABG mortality among the higher-than-expected hospitals declined from 7.12 percent to 2.77 percent, a 61 percent decrease. Mortality rate performance in the other two groups improved as well: from 4.24 percent in 1989 to 2.51 percent in 1992 among as-expected hospitals, a 41 percent reduction; and from 2.72 percent to 2.19 percent, a 19 percent reduction, among the lower-than-expected group of hospitals. Hannan and colleagues hypothesized that observed improvements in CABG were at least partially attributable to quality improvements resulting from public release of performance data. Hannan et al. (1994), in *New York's Cardiac Surgery*, considered three alternative explanations for the findings observed:

- Mortality improvements in New York could simply be reflective of broader national patterns.
- Improvements could reflect intentional "upcoding" of patient severity by hospitals to increase expected number of CABG deaths and thereby to reduce risk-adjusted mortality rates.
- Improvements in New York CABG mortality rates could be the result of selection bias resulting from surgeons' refusal to accept high-risk patients.

Hannan et al. (1990) concluded that the study's findings were unlikely to have been influenced by these factors. First, they noted that comparison of New York CABG mortality rates with those from two other regions in the Northeast indicated that New York's rates were lower than elsewhere. They also observed that a medical record audit of 10 hospitals conducted by the New York Department of Health revealed no systematic patterns of coding errors. In response to the final concern, Hannan et al. (1994), in *New York's Cardiac Surgery*,

presented data indicating that risky patients were not being turned away. During the first 4 years of the data release, CABG surgery volume in New York hospitals increased 31 percent, and the "riskiness" of patients undergoing CABG procedures, measured by the expected mortality rate, increased 25 percent during the period.

Other researchers, however, have recently begun to dispute these conclusions. For example, Ghali et al. (1997) present data showing that in Massachusetts, a state in which mortality rate data are not published, risk-adjusted CABG mortality declined from 1990 to 1994 at about the same rate as that documented in Hannan et al. (1994), in *Improving the Outcomes*, for the equivalent period. Green and Wintfeld (1995) also criticized the conclusions of Hannan et al. (1994), in *New York's Cardiac Surgery*, for other reasons. One point cited by Green and Wintfeld (1995) was that, contrary to NYSDOH medical record audit findings, prevalence of risk factors included in the CSRS model (renal failure, CHF, chronic obstructive pulmonary disease [COPD], unstable angina, and low ejection fraction) showed large, sudden increases in prevalence after the initial public release of mortality data. From 1989 to 1991, prevalence of renal failure among New York CABG patients went from 0.4 percent to 2.8 percent, prevalence of CHF went from 1.7 percent to 7.6 percent, and prevalence of COPD went from 6.9 percent to 17.4 percent. At one hospital, prevalence of COPD went from 1.8 percent to 52.9 percent, and at another hospital, prevalence of unstable angina went from 1.9 percent to 20.8 percent (Green and Wintfeld 1995). Omoigui et al. (1996) have charged that improvements in New York CABG mortality rates are at least partially due to outmigration of the riskiest cases to other states, especially to Ohio, home of the Cleveland Clinic. The accuracy of this claim is supported by evidence from Schneider and Epstein (1996). Since 1992, the state of Pennsylvania has produced a CABG mortality report similar to that of New York. In a survey of Pennsylvania cardiologists, Schneider and Epstein (1996) found that 59 percent of respondents reported experiencing increased difficulty in finding surgeons willing to perform CABG surgery on very risky patients since implementation of the report. Of the cardiac surgeons responding to the survey, 63 percent indicated that they were less willing to operate on very risky patients because of the report.

Accuracy of conclusions from Rosenthal, Quinn, and Harper (1997) and studies by Hannan and colleagues are challenged on other grounds as well. Patterns of improvement among mortality rate terciles, discussed by Hannan et al. (1994), in *New York's Cardiac Surgery*, almost certainly reflect the influence of regression to the mean. In fact, each of the studies is subject to virtually all of the confounding factors listed by Campbell and Stanley (1963) as threats to validity of quasi-experimental design studies. Because there were no con-

trols, it is plausible that declines in mortality observed for pneumonia and CHF cases in northeastern Ohio and those for CABG cases in New York—as Ghali et al. (1997) note—were attributable solely to general improvements in clinical technology.

SUMMARY AND CONCLUSION

Methodologies for risk-adjusting outcomes are now widely available, and evaluations of individual hospital performance in terms of mortality rates and other outcomes are being produced and disseminated with increasing regularity by state agencies, business coalitions, hospital associations, and commercial vendors of health data. Despite the amount of resources devoted to collecting, analyzing, and disseminating data, and despite continuing criticisms of data accuracy, little attention has been given to questions about the validity of quality performance information made available to the public.

The purpose of this article was to review research findings related to the question of whether hospitals' risk-adjusted mortality rates can be considered valid as indicators of quality of care. Our conclusions are as follows:

EMPIRICAL STUDIES OF HOSPITAL-LEVEL PROCESS-OUTCOME RELATIONSHIPS

The three studies that used regression or correlation analyses to investigate relationships between hospitals' risk-adjusted mortality rates and process-based indicators of quality produced positive, but equivocal, results. Although Williams (1979) found evidence of relationships between hospitals' risk-adjusted mortality rates and quality-related structural and process characteristics, the findings are difficult to generalize because of the stepwise methodology used. For cardiac surgery, AMI, and pneumonia cases, Thomas (1991) found significant hospital-level correlations between risk-adjusted mortality rates and quality, but no significant relationships were found for seven other conditions studied. While Hartz et al. (1993) found significant positive correlations between hospitals' overall risk-adjusted Medicare mortality rates and rates of peer review quality problems in 14 states, they found no similar relationships in 24 other states studied.

Findings from the five predictive validity studies reviewed in this article are similarly equivocal. Knaus et al. (1986) noted important quality-related clinical management differences between low-outlier and high-outlier ICUs; but only 13 ICUs participated in this study, and judgments about unit quality performance were post hoc. DuBois et al. (1987) reported finding significant differences between preventable death rates of pneumonia patients treated in

low-outlier and high-outlier hospitals. However, the finding is suspect because of the low reliability of preventable-death judgments by reviewers. Also, DuBois et al. (1987) found no differences in preventable death rates for AMI and stroke patients, and no differences among hospitals for any of the conditions studied when quality was judged using explicit, condition-specific criteria. Park et al. (1990) found that average-quality scores for CHF cases treated in high-outlier hospitals were significantly lower than scores for patients treated in low-outlier hospitals. However, quality score differences for AMI cases were not significant. Although a review of medical records in New York (Hannan et al. 1990) showed a tenfold difference in quality problem rates between high–mortality rate and low–mortality rate outlier hospitals, the study involved only 63 patients, and the reported quality problem rate for low-outlier hospitals (4.35 percent) represented only one case. The most extensive empirical validation to date of risk-adjusted mortality rates as indicators of hospital quality is the study conducted by OSHPD (1996). Significant differences were observed between low-outlier and high-outlier hospitals in rates of compliance, with six explicit processes believed to relate to good-quality care. However, rates of compliance with 12 other criteria were not different between the samples of hospitals.

STUDIES INVESTIGATING CONDITIONS THAT AFFECT DEGREE OF VALIDITY OF MORTALITY RATE MEASURES

For mortality rates to be valid as indicators of hospital quality performance, it is necessary that a population-level relationship exists between quality of care and patient mortality risk. The most credible evidence of such a relationship was reported by Kahn et al. (1990), in *Measuring Quality*, and by Rubenstein et al. (1990). Using a highly reliable medical record review methodology, Rubenstein et al. (1990) determined that cases that had received poor-quality care were more than twice as likely to die than other cases. In a study of relationships between quality of care and patient mortality rates with Medicare peer review data, Thomas (1991) found poor quality to be associated with significantly higher mortality risk for patients admitted for cardiac surgery, AMI, cardiac arrhythmia, pneumonia, and femur/pelvic fractures. No quality-mortality relationships were observed for five other diagnostic groups studied.

Since the mid-1980s, when public reports of hospital mortality rates first began to appear, the greatest concern of both advocates and critics has been the adequacy of risk adjustments used to control for systematic errors in hospital mortality rates. In 1994, Iezzoni and colleagues summarized extensive literature on this issue (Iezzoni 1994). Since then, these researchers conducted addi-

tional studies to investigate the reliability of mortality risk assessments. They found, as summarized by Iezzoni (1997), that identification of hospitals as either low–mortality rate outliers or high–mortality rate outliers is quite sensitive to the specific methodology chosen to adjust for casemix and severity differences among hospitals. Because many of the risk-adjustment methodologies evaluated by Iezzoni and colleagues are currently used in hospital report cards, it must follow that at least some published measures of hospital mortality rate performance are inaccurate. Although substantial progress has been made in the past decade in risk-adjustment technology, evidence suggests that risk-adjusted mortality rates are still subject to systematic measurement error.

In 1990, Park et al. provided evidence that observed hospital mortality rates included significant amounts (56 percent to 82 percent) of random measurement error. With multiple years of data on CABG mortality in California hospitals, Luft and Romano (1993) showed that 2 years after identification as mortality rate outliers, mortality rates among high-outlier hospitals averaged 31 percent higher than expected, and among low–mortality rate outliers 28 percent lower than expected. However, Luft and Romano (1993) noted that rates varied significantly among individual facilities and that no hospital identified as an outlier with 1987 data remained an outlier in 1989. Other recent articles have focused on the influence of random measurement error on accuracy of identifying individual poor-quality hospitals. Findings from these studies (Hofer and Hayward 1996; Zalkind 1997; Thomas and Hofer forthcoming) suggest that, as an indicator of hospital quality performance, risk-adjusted mortality rates are both insensitive and nonspecific. On average, even with perfect risk adjustment, the indicator fails to identify 88 percent of the facilities that deliver poor-quality care. Furthermore, 60 percent of high–mortality rate outliers are falsely marked as poor-quality providers.

STUDIES IN WHICH VALIDITY OF MORTALITY RATE IS INFERRED FROM OBSERVED BEHAVIORAL RESPONSES

Each of the studies reviewed in this section presented evidence that hospital mortality rates declined over time in response to publication of mortality rate performance data (Hannan et al., *Improving the Outcomes*, 1994; Rosenthal, Quinn, and Harper et al. 1997). In critiquing these articles, we noted that experimental design problems make both studies susceptible to virtually all of the confounding factors identified by Campbell and Stanley (1963). While raising interesting issues and stimulating constructive debate, findings from these studies cannot be viewed as credible support for the validity of risk-adjusted mortality rates as indicators of hospital quality-of-care performance.

It is interesting to note that when identifying poor-quality hospitals on the basis of risk-adjusted mortality rates, problems with systematic measurement error and random measurement error can compound each other. Hofer and Hayward (1996) found that with hospital mortality rates calculated on the basis of all medical admissions, the measure's sensitivity for identifying poor-quality hospitals was 35 percent and its predictive error was 48 percent. However, when mortality rates were calculated separately for specific medical diagnoses, due to effects of smaller patient sample sizes, sensitivity dropped to under 10 percent and predictive error increased to 75 percent. As a strategy for increasing sample sizes and thereby reducing random measurement error in hospital performance reports, some organizations now determine mortality rates for clusters of related diagnoses. For example, one report (Michigan Hospital Association 1997) lists risk-adjusted mortality rates for a combined group of medical diagnoses, including admissions for CHF, stroke, pneumonia, chronic lung disease, and gastrointestinal bleeding; and for a combined group of surgical admissions, including lung surgery, lower bowel surgery, spine surgery, prostate surgery, and hysterectomy. However, while potentially reducing random measurement error, this approach simultaneously increases *systematic* measurement error, since it involves combining data on several measures, each subject to some degree of measurement error, in proportions that differ from hospital to hospital. Although the increase in systematic measurement error associated with clustering diagnoses might be minimized through indirect standardization of diagnostic casemix across hospitals, to our knowledge, this approach has not been used in hospital performance reports.

In summary, our review suggests that patients who receive poor-quality hospital care experience elevated mortality risks. We find also that, on average, patients treated in hospitals identified as high–mortality rate outliers may be at somewhat greater risk of receiving poor-quality care than patients treated in nonoutlier hospitals. Nevertheless, our principal finding supports the recommendation of Sisk et al. (1990) for caution in using mortality rate as an indicator of quality. Indeed, we feel compelled to make the statement even stronger. At the beginning of this article, we posed three questions that can now be answered quite simply.

1. Do hospitals that deliver poor-quality care experience higher risk-adjusted mortality rates than other hospitals?
   Yes.
2. Can we be confident that hospitals whose mortality rates are significantly higher than expected are actually poor-quality providers?
   No.

3. For any two hospitals, is a difference in risk-adjusted mortality rates indicative of a difference in quality performance?
   No.

As shown here, evidence that supports the validity of mortality rates as measures of hospital quality is quite fragile, while evidence is quite strong that hospital mortality rates are subject to significant measurement error. Because the number of published hospital performance reports continues to increase each year, it appears that purchasers, data vendors, and many researchers continue to overestimate the ability of available methodologies to adjust accurately for differences in hospitals' patient populations. It is also clear that they do not adequately comprehend the impact of random measurement error on the validity of mortality rate-based quality inferences. Jencks (1995) states that data for public release "almost always require higher standards of rigor and proof" than data used internally, for example, for quality improvement studies. Because hospital mortality rate statistics do not meet even minimal standards for publication, mortality data reports do more to misinform than to inform the public.

NOTES

1. A subset of findings from this report is published in Thomas, Holloway, and Guire (1993).
2. We put *necessary* in quotes because the usual use of this term implies binary application—conditions are either met or they are not met. McAuliffe (1984), quoted above, notes that since no measure of an abstract concept can be perfectly valid, validity must be considered a matter of degree. The conditions that we identify in this section influence the *degree* of validity possessed by a measure; they do not determine the presence or absence of validity.
3. This negative finding was considered partially attributable to the small number of patient deaths in this diagnosis category.

REFERENCES

Berwick, D. M., and D. L. Wald. 1990. Hospital Leaders' Opinions of the HCFA Mortality Data. *Journal of the American Medical Association* 263:247-49.

Blumberg, M. S. 1987. Comments on HCFA Hospital Death Rate Statistical Outliers. *Health Services Research* 21:715-39.

———. 1991. Biased Estimates of Expected Acute Myocardial Infarction Mortality Using MedisGroups Admission Severity Groups. *Journal of the American Medical Association* 265:2965-70.

Brook, R. H., A. Davies-Avery, S. Greenfield, L. J. Harris, R. Lelah, N. E. Solomon, and J. E. Ware. 1977. Assessing the Quality of Medical Care Using Outcome Measures: An Overview of the Method. *Medical Care* 15 (9, Supp.).

California Office of Statewide Health Planning and Development (COSHPD). 1996. *Report of the California Hospital Outcomes Project: Hospital Specific Detailed Statistical Tables*. Vol. 2, *Technical Appendix*. Sacramento, CA: OSHPD.

Campbell, D. T., and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Daley, J. 1994. Validity of Risk-Adjustment Methods. In *Risk Adjustment for Measuring Health Care Outcomes*, ed. L. I. Iezzoni; 239-62. Ann Arbor, MI: Health Administration Press.

DesHarnais, S., J. D. Chesney, R. Wroblewski, S. T. Fleming, and L. F. McMahon. 1988. The Risk-Adjusted Mortality Index. A New Measure of Hospital Performance. *Medical Care* 26:1129-48.

Donabedian, A. 1988. Quality Assessment and Assurance: Unity of Purpose, Diversity of Means. *Inquiry* 25 (1): 173-92.

DuBois, R. W. 1989. Hospital Mortality as an Indicator of Quality. In *Providing Quality Care: The Challenge to Clinicians*, eds. N. Goldfield and D. B. Nash; 107-31. Philadelphia: American College of Physicians.

DuBois, R. W., W. H. Rogers, J. H. Moxley, D. Draper, and R. H. Brook. 1987. Hospital Inpatient Mortality: Is It a Predictor of Quality? *New England Journal of Medicine* 317:1674-80.

Ghali, W. A., A. S. Ash, R. E. Hall, and M. A. Moskowitz. 1997. Statewide Quality Improvement Initiatives and Mortality after Cardiac Surgery. *Journal of the American Medical Association* 277:379-82.

Goldman, R. L. 1992. The Reliability of Peer Assessments of Quality of Care. *Journal of the American Medical Association* 267:958-60.

Goss, M.E.W., and J. I. Reed. 1974. Evaluating the Quality of Hospital Care through Severity-Adjusted Death Rates: Some Pitfalls. *Medical Care* 12:202-13.

Green, J., and N. Wintfeld. 1995. Report Card on Cardiac Surgeons: Assessing New York State's Approach. *New England Journal of Medicine* 332:1229-32.

Green, J., N. Wintfeld, M. Krasner, and C. Wells. 1997. In Search of America's Best Hospitals: The Promise and Reality of Quality Assessment. *Journal of the American Medical Association* 277:1152-55.

Hadorn, D. C., E. B. Keeler, W. H. Rogers, and R. H. Brook. 1993. *Assessing the Performance of Mortality Prediction Models*. Rand Report ISBN: 0-8330-1335-1. Santa Monica, CA: RAND.

Hannan, E. L., H. Kilburn Jr., M. L. Lindsey, and R. Lewis. 1992. Clinical versus Administrative Data Bases for CABG Surgery: Does It Matter? *Medical Care* 30:892-907.

Hannan, E. L., H. Kilburn Jr., J. F. O'Donnell, G. Lukacik, and E. P. Shields. 1990. Adult Open Heart Surgery in New York State. *Journal of the American Medical Association* 264:2768-74.

Hannan, E. L., H. Kilburn Jr., M. Racz, E. Shields, and M. R. Chassin. 1994. Improving the Outcomes of Coronary Artery Bypass Surgery in New York State. *JAMA* 271:761-66.

Hannan, E. L., D. Kumar, M. Racz, A. Siu, and M. R. Chassin. 1994. New York's Cardiac Surgery Reporting System: Four Years Later. *Annals of Thoracic Surgery* 58:1852-57.

Hannan, E. L., M. J. Racz, J. G. Jollis, and E. D. Peterson. 1997. Using Medicare Claims

Data to Assess Provider Quality for CABG Surgery: Does It Work Well Enough? *Health Services Research* 31:569-678.

Hannan, E. L., and A. A. Yazici. 1988. Critique of the 1987 HCFA Mortality Study Based on New York Data. Albany: New York State Department of Health.

Hartz, A. J., M. S. Gottlieb, E. M. Kuhn, and A. A. Rimm. 1993. The Relationship between Adjusted Hospital Mortality and the Results of Peer Review. *Health Services Research* 27:767-77.

Hartz, A. J., and E. M. Kuhn. 1994. Comparing Hospitals That Perform Coronary Artery Bypass Surgery: The Effect of Outcome Measures and Data Sources. *American Journal of Public Health* 84:1609-14.

Hill, C. A., K. L. Winfrey, and B. A. Rudolph. 1997. "Best Hospitals:" A Description of the Methodology for the Index of Hospital Quality. *Inquiry* 34:80-90.

Hofer, T. P., and R. A. Hayward. 1996. Identifying Poor-Quality Hospitals: Can Hospital Mortality Rates Detect Quality Problems for Medical Diagnoses? *Medical Care* 34:737-53.

Iezzoni, L. I., ed. 1994. *Risk Adjustment for Measuring Health Care Outcomes.* Ann Arbor, MI: Health Administration Press.

———. 1997. The Risks of Risk Adjustment. *Journal of the American Medical Association* 278:1600-1607.

Iezzoni, L. I., A. S. Ash, M. Shwartz, J. Daley, J. S. Hughes, and Y. D. Mackiernan. 1995. Predicting Who Dies Depends on How Severity Is Measured: Implications for Evaluating Patient Outcomes. *Annals of Internal Medicine* 123:763-770.

Iezzoni, L. I., M. Shwartz, A. S. Ash, J. S. Hughes, J. Daley, and Y. D. Mackiernan. 1995. Using Severity-Adjusted Stroke Mortality Rates to Judge Hospitals. *International Journal for Quality in Health Care* 7 (2): 81-94.

———. 1996. Severity Measurement Methods and Judging Hospital Death Rates for Pneumonia. *Medical Care* 34:11-28.

Institute of Medicine (IOM). 1974. *Advancing the Quality of Health Care: Key Issues and Fundamental Principles. A Policy Statement by a Committee of the Institute of Medicine.* Washington, DC: National Academy of Sciences.

Jacobs, D. M., and N. D. Jacobs. 1974. *The PEP Primer: The JCAH Performance Evaluation Procedure for Auditing and Improving Physician Care.* Chicago: Quality Review Center, Joint Commission on Accreditation of Hospitals.

Jencks, S. F. 1995. Measuring Quality of Care under Medicare and Medicaid. *Health Care Financing Review* 16 (4): 39-54.

Jencks, S. F., D. K. Williams, and T. L. Kay. 1988. Assessing Hospital-Associated Deaths from Discharge Data: The Role of Length of Stay and Comorbidities. *Journal of the American Medical Association* 260:2240.

Kahn, K. L., E. B. Keeler, M. J. Sherwood, W. H. Rodgers, et al. 1990. Comparing Outcomes of Care before and after Implementation of the DRG-Based Prospective Payment System. *Journal of the American Medical Association* 264:1984-88.

Kahn, K. L., W. H. Rogers, L. V. Rubenstein, M. J. Sherwood, E. J. Reinisch, E. B. Keeler, D. Draper, J. Kosecoff, and R. H. Brook. 1990. Measuring Quality of Care with Explicit Process Criteria before and after Implementation of the DRG-Based Prospec-

tive Payment System. *Journal of the American Medical Association* 264:1969-73.

Keeler, E. B., K. L. Kahn, D. Draper, M. J. Sherwood, L. V. Rubenstein, E. J. Reinisch, J. Kosecoff, and R. H. Brook. 1990. Changes in Sickness at Admission Following the Introduction of the Prospective Payment System. *Journal of the American Medical Association* 264:1962-68.

Knaus, W. A., E. A. Draper, D. P. Wagner, and J. E. Zimmerman. 1986. An Evaluation of Outcome from Intensive Care in Major Medical Centers. *Annals of Internal Medicine* 104:410-418.

Knaus, W. A., D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, and A. Damiano. 1991. The APACHE III Prognostic System: Risk Prediction of Hospital Mortality for Critically Ill Hospitalized Adults. *Chest* 100:1619.

Landon, B., Iezoni, L. I., A. S. Ash, M. Shwartz, J. Daley, J. S. Hughes, and Y. D. Mackiernan. 1996. Judging Hospitals by Severity-Adjusted Mortality Rates: The Case of CABG Surgery. *Inquiry* 33:155-66.

Lave, J. R., and L. B. Lave. 1971. The Extent of Role Differentiation among Hospitals. *Health Services Research* 6:15-38.

Luft, H. S., and P. S. Romano. 1993. Chance, Continuity, and Change in Hospital Mortality Rates. *Journal of the American Medical Association* 270:331-37.

McAuliffe, W. E. 1984. A Validation Theory for Quality Assessment. In *Hospital Quality Assurance: Risk Management and Program Evaluation*, eds. J. J. Pena, A. N. Haffner, B. Rosen, and D. W. Light; 157-174. Gaithersburg, MD: Aspen.

McClure, W. 1985. Buying Right: How to Do It. *Business & Health* (October): 41-44.

Michigan Hospital Association. 1997. *Michigan Hospital Performance Report*. Vol. 2. Lansing: Michigan Hospital Association.

New York State Department of Health, Office of Health Systems Management, Bureau of Healthcare Research (NYSDOH). *Investigation of Quality of Care in Hospitals*. Albany, NY: New York State Department of Health.

Omoigui, N. A., D. P. Miller, K. J. Brown, K. Annan, D. Cosgrove, B. Lytle, F. Loop, and E. J. Topol. 1996. Outmigration for Coronary Bypass Surgery in an Era of Public Dissemination of Clinical Outcomes. *Circulation* 93:27-33.

Park, R. E., R. H. Brook, J. Kosecoff, J. Keesey, L. Rubenstein, E. Keeler, K. L. Kahn, W. H. Rogers, M. R. Chassin, et al. 1990. Explaining Variations in Hospital Death Rates: Randomness, Severity of Illness, Quality of Care. *Journal of the American Medical Association* 264:484-90.

Roemer, M. I., A. T. Moustafa, and C. E. Hopkins. 1968. A Proposed Hospital Quality Index: Hospital Death Rates Adjusted for Case Severity. *Health Services Research* 3:96-118.

Romano, P. S., A. Zach, H. S. Luft, J. Rainwater, L. L. Remy, and D. Campa. 1995. The California Hospital Outcomes Project: Using Administrative Data to Compare Hospital Performance. *Joint Commission Journal on Quality Improvement* 21:668-82.

Rosenthal, G. E., and D. L. Harper. 1994. Cleveland Health Quality Choice: A Model for Community-Based Outcomes Assessment. *Joint Commission Journal on Quality Improvement* 20:425-44.

Rosenthal, G. E., L. Quinn, and D. L. Harper. 1997. Declines in Hospital Mortality Associated with a Regional Initiative to Measure Hospital Performance. *American Journal of Medical Quality* 15:103-12.

Rubenstein, L. V., K. L. Kahn, E. J. Reinisch, M. J. Sherwood, W. H. Rogers, C. Kamberg, D. Draper, and R. H. Brook. 1990. Changes in Quality of Care for Five Diseases Measured by Implicit Review, 1981-1986. *Journal of the American Medical Association* 264:1974-79.

Rubin, H. R., W. H. Rogers, K. L. Kahn, L. V. Rubenstein, and R. H. Brook. 1992. Watching the Doctor Watchers: How Well Do Peer Review Organization Methods Detect Hospital Quality Problems? *Journal of the American Medical Association* 267:2349-54.

Schneider, E. C., and A. M. Epstein. 1996. Influence of Cardiac-Surgery Performance Reports on Referral Practices and Access to Care. *New England Journal of Medicine* 335:251-56.

Sisk, J. E., D. M. Dougherty, P. M. Ehrenhaft, G. Ruby, and B. A. Mitchner. 1990. Assessing Information for Consumers on the Quality of Medical Care. *Inquiry* 27:263-72.

Steen, P. M., A. C. Brewster, R. C. Bradbury, E. Estabrook, and J. A. Young. 1993. Predicted Probabilities of Hospital Death as a Measure of Admission Severity of Illness. *Inquiry* 30:128.

Thomas, J. W. 1991. *Validating Risk-Adjusted Outcomes as Measures of Quality of Care in Hospitals*. Report prepared for the Minnesota Coalition on Health. Ann Arbor: University of Michigan, School of Public Health, Department of Health Services Management and Policy.

Thomas, J. W., and T. P. Hofer. Forthcoming. Accuracy of Risk-Adjusted Mortality Rate as a Measure of Hospital Quality of Care. *Medical Care*.

Thomas, J. W., J. J. Holloway, and K. E. Guire. 1993. Validating Risk-Adjusted Mortality as an Indicator for Quality of Care. *Inquiry* 30:6-22.

U.S. Congress, General Accounting Office (USCGAO). 1988. *Medicare: Improved Patient Outcome Analyses Could Enhance Quality Assessment*. Washington, DC: Government Printing Office.

Vladeck, B. C., E. J. Goodwin, L. P. Myers, and M. Sinsi. 1988. Consumers and Hospital Use: The HCFA "Death List." *Health Affairs* 7:122-25.

Williams, R. L. 1979. Measuring the Effectiveness of Perinatal Medical Care. *Medical Care* 17:95-110.

Young, W. W., S. Kohler, and J. Kowalski. 1994. PMC Patient Severity Scale: Derivation and Validation. *Health Services Research* 29:367.

Zalkind, D. L. 1997. Mortality Rates as an Indicator of Hospital Quality. *Hospital & Health Services Administration* 42:3-15.