Judging Hospitals by Severity-Adjusted Mortality Rates: The Influence of the Severity-Adjustment Method

Lisa I. Iezzoni, MD, MSc, Arlene S. Ash, PhD, Michael Shwartz, PhD, Jennifer Daley, MD, John S. Hughes, MD, and Yevgenia D. Mackiernan

ABSTRACT

Objectives. This research examined whether judgments about a hospital's risk-adjusted mortality performance are affected by the severity-adjustment method.

Methods. Data came from 100 acute care hospitals nationwide and 11 880 adults admitted in 1991 for acute myocardial infarction. Ten severity measures were used in separate multivariable logistic models predicting in-hospital death. Observed-to-expected death rates and *z* scores were calculated with each severity measure for each hospital.

Results. Unadjusted mortality rates for the 100 hospitals ranged from 4.8% to 26.4%. For 32 hospitals, observed mortality rates differed significantly from expected rates for 1 or more, but not for all 10, severity measures. Agreement between pairs of severity measures on whether hospitals were flagged as statistical mortality outliers ranged from fair to good. Severity measures based on medical records frequently disagreed with measures based on discharge abstracts.

Conclusions. Although the 10 severity measures agreed about relative hospital performance more often than would be expected by chance, assessments of individual hospital mortality rates varied by different severity-adjustment methods. (*Am J Public Health.* 1996;86:1379–1387)

Introduction

Report cards on hospitals and physicians are increasingly used to compare provider performance along a variety of dimensions, including patient outcomes and costs.[1–6] Managed care companies and business coalitions use these reports to select preferred providers for networking and contracting.[7,8] States, including California, Florida, New York, and Pennsylvania, generate and publicly disseminate hospital-specific comparisons of total charges and patient death rates, ostensibly to assist consumers to choose among hospitals.[3–6] Therefore, especially in highly competitive regions, methods employed to judge performance are critically important to health care providers and could substantially affect local health care delivery systems.[1–8]

In most instances, performance reports are touted as indicators of relative provider quality. Despite these claims, a recent US General Accounting Office review of governmental and private payer report card initiatives noted that few valid hospital quality measures exist and that current databases rarely provide insight into important patient outcomes.[3] For example, although hospital mortality rates are often a centerpiece of such report cards, the relationship of death rates to quality of care remains controversial and unproven. The report also emphasized that unless findings are adjusted for patient characteristics, "conclusions about quality based on an evaluation of [patient] outcomes might be erroneous."[3(p.4)]

Because some facilities treat sicker patients than others, hospital comparisons need to control for patient risk.[9,10] However, methods for quantifying patient risk are also controversial. Since the early 1980s, numerous severity measures have been developed specifically for comparing hospitals or large patient groups.[11–13] Many are commercial, proprietary products, marketed to hospitals, government officials, legislators, and business leaders. Some locales and payers require hospitals to report information using specific severity measures[3–7]; for example, since 1986, Pennsylvania hospitals have had to report inpatient severity of illness using MedisGroups. Choices of severity measure are often idiosyncratic, sometimes based primarily on the vendor's marketing approach. Articles, usually by the developers of the measures, describe individual severity measures,[14–31] but few studies by independent investigators involving multiple severity measures have been reported,[32–34] and fewer have examined the effect of different severity measures on hospital comparisons.[35,36]

Given the potential impact of these measures on providers and patients, comprehensive evaluations of their validity

Lisa I. Iezzoni, Jennifer Daley, and Yevgenia D. Mackiernan are with the Division of General Medicine and Primary Care, Harvard Medical School, Beth Israel Hospital, the Charles A. Dana Research Institute, and the Harvard–Thorndike Laboratory, Boston, Mass. Arlene S. Ash is with the Health Care Research Unit, Boston University Medical Center, Boston. Michael Shwartz is with the Health Care Management Program and Operations, Boston University. John S. Hughes is with the Department of Medicine, West Haven Veterans Affairs Medical Center, West Haven, Conn.

Requests for reprints should be sent to Lisa I. Iezzoni, MD, MSc, Division of General Medicine and Primary Care, Department of Medicine, Beth Israel Hospital, 330 Brookline Ave, Boston, MA 02215.

This paper was accepted January 18, 1996.

Note. The views expressed in this paper are solely those of the authors.

TABLE 1—Description of 10 Severity-Adjustment Methods for Hospital Mortality Rates

System	Source/Vendor	Data Used and Definition of Severity	Classification Approach
Clinical data–based methods			
MedisGroups (i.e., Atlas MQ)	MediQual Systems, Inc, Westborough, Mass	Clinical variables	
Original version[21–23]		Clinical instability indicated by in-hospital death	Admission score 0, 1, 2, 3, or 4
Empirical version[24]		In-hospital death	Probability ranging from 0 to 1
Physiology Score 1	Patterned after the Acute Physiology Score (APS) of the Acute Physiology and Chronic Health Evaluation (APACHE), version II[26,27]	12 clinical variables; in-hospital mortality for patients in intensive care unit	Integer score starting with 0; APACHE II's APS ranges from 0 to 60
Physiology Score 2	Patterned after APS of APACHE, version III[28,29]	17 clinical variables; in-hospital mortality for patients in intensive care unit	Integer score starting with 0; APACHE III's APS ranges from 0 to 252
Discharge abstract–based methods: methods with a clinical definition of severity			
Disease Staging[18–20]	SysteMetrics/MEDSTAT Group, Santa Barbara, Calif	Discharge abstract	
Mortality probability		Probability of in-hospital death	Probability ranging from 0 to 1
Stage		Stage of disease based on risk of death or functional impairment	Three stages (1.0, 2.0, and 3.0) with substages within each stage
Patient Management Categories (PMCs)—Severity Score[25]	Pittsburgh Research Institute, Pittsburgh, Pa	Discharge abstract; in-hospital morbidity and mortality	Score of 1, 2, 3, 4, 5, 6, or 7
Comorbidity Index	Developed by Charlson et al.[16]; coded version patterned after Deyo et al.[17]	Discharge abstract; risk of death within 1 year of medical hospitalization	Integer from additive scale representing number and severity of comorbidities
Discharge abstract–based methods: methods with a resource-based definition of severity			
All Patient Refined Diagnosis-Related Groups (APR-DRGs)[14,15]	3M Health Information Systems, Wallingford, Conn	Discharge abstract; total hospital charges	Four severity classes (A, B, C, D) within adjacent DRGs[a]
Refined Diagnosis-Related Groups (R-DRGs)[15,30,31]	Yale University refinement of DRGs provided by Yale Project Director Karen Schneider, Health Systems Consultants, New Haven, Conn	Discharge abstract; length of hospital stay, total hospital charges	Three severity classes (B, C, D) within adjacent medical DRGs[a]; "early" deaths grouped in lowest severity class

[a]Adjacent DRGs are formed by grouping individual DRGs previously split by complications and comorbidities.

are needed. Here, we consider a single question reflecting one way these measures are used around the country. We employed "off-the-shelf" severity measures (Table 1) to construct risk-adjusted predictions of in-hospital deaths and asked the following question: does choice of severity measure affect which hospitals are seen as having lower- or higher-than-expected death rates?

Methods

Severity Measures

We examined 10 measures (Table 1) representative of approaches used in comparing hospital mortality rates around the country[3–7] (we did not consider measures developed primarily for clinical or health services research). The two MedisGroups and two physiology scores use clinical data abstracted from medical records. Six measures rate patients with standard data from hospital discharge abstracts,[37–39] such as age, sex, and diagnoses and procedures coded with the *International Classification of Diseases,* 9th Revision, Clinical Modification (ICD-9 CM). The discharge abstract-based measures have either clinical or resource-based definitions of severity (Table 1). Even resource-derived measures are sometimes employed to look at mortality,[40] although these uses are rarely published in scholarly settings (e.g., they are conducted by proprietary health care information companies who sell results to hospitals). Measures assign either numerical scores or values on a continuous scale (Table 1).

Database

To assign severity scores, computerized algorithms were applied to a data file extracted from the 1992 MedisGroups Comparative Database, containing the clinical information collected on hospitalized patients with MedisGroups' data

TABLE 2—Examples of Relative Mortality Performance for Five Hospitals: Ranks by Unadjusted Death Rates and by z Scores Associated with Observed-to-Expected Death Rates Calculated by Different Severity Methods

	Hospital				
	A	B	C	D	E
No. died/total no. cases	22/204	38/246	24/130	20/106	43/224
Death rate, %	10.8	15.4	18.5	18.9	19.2
Decile rank by unadjusted death rate[a]	3	8	9	9	10
z score (decile rank by *z* score)[b]					
MedisGroups—Original	−2.82 (1)	−1.14 (2)	1.83 (10)	0.53 (7)	2.26 (10)
MedisGroups—Empirical	−2.04 (1)	−2.55 (1)	1.75 (10)	0.95 (8)	2.66 (10)
Physiology Score 1	−3.05 (1)	−1.99 (1)	2.38 (10)	1.19 (8)	2.70 (10)
Physiology Score 2	−2.97 (1)	−1.79 (2)	2.56 (10)	0.23 (6)	3.15 (10)
Disease Staging—PR	−1.04 (3)	−2.54 (1)	0.88 (7)	3.08 (10)	1.32 (9)
Disease Staging—stage	−1.13 (3)	−1.75 (1)	1.29 (9)	1.71 (9)	1.54 (9)
PMCs—Severity Score	−2.27 (1)	−1.90 (1)	0.67 (7)	2.86 (10)	1.12 (8)
Comorbidity Index	−1.72 (1)	−0.50 (4)	1.71 (10)	1.49 (9)	2.20 (10)
APR-DRGs	−2.13 (1)	−3.15 (1)	1.05 (8)	2.12 (10)	1.73 (9)
R-DRG	−1.67 (1)	0.39 (7)	1.45 (9)	1.17 (9)	3.41 (10)

Note. APR-DRGs = All Patient Refined Diagnosis-Related Groups; Disease Staging—PR = Disease Staging mortality probability; PMCs—Severity Score = Patient Management Categories—Severity Score; R-DRG = Refined Diagnosis-Related Groups.
[a]Decile of rank of hospital by actual death rate, unadjusted for age, sex, or patient severity of illness. 1 = death rate in the lowest 10%; 10 = death rate in the highest 10%.
[b]Decile of rank of *z* score. 1 = *z* score in the lowest 10%; 10 = *z* score in the highest 10%.

collection protocol.[41,42] Hospitals using MedisGroups provide these data to its vendor, MediQual Systems. The 1992 MedisGroups Comparative Database is a subset of a larger file, containing all calendar year 1991 discharges from 108 acute care hospitals chosen by MediQual Systems for good data quality and variety of characteristics.

Because of our interest in hospital-level analyses, we eliminated eight institutions with fewer than 30 eligible cases (83 patients total). Hospital characteristics were taken from the American Hospital Association annual survey.

MediQual Systems provided original and empirical admission MedisGroups scores; other scores had to be assigned. The MedisGroups database includes values of key clinical findings from the admission period (generally the first 2 hospital days) abstracted from medical records during MedisGroups reviews.[21–24] We used key clinical findings to create physiology scores patterned after the Acute Physiology and Chronic Health Evaluation, Version II (APACHE II) and APACHE III by assigning weights specified by APACHE II and III (e.g., a pulse of 145 beats/minute had a weight of 13 points for APACHE III[28]) and summing these weights to produce scores. We could not exactly replicate APACHE II or III Acute Physiology Scores because complete values for required physiologic variables were unavailable: MedisGroups truncated data collection in broadly defined normal ranges.[43] Previous research demonstrated that these derived physiology scores performed well compared with actual APACHE II scores.[43]

The MedisGroups database also contains standard discharge abstract information listed by hospitals, including up to 20 ICD-9 CM discharge diagnosis codes and 50 ICD-9 CM procedure codes. We assigned a code-based version of the Charlson comorbidity index,[16] using an approach adapted from Deyo et al.[17] Other severity scoring was performed by the vendors (Table 1). Based on their specifications, we supplied computer files containing necessary discharge abstract data elements from the MedisGroups database. Vendors applied their severity software and returned the data to us after scoring.

Study Sample and Outcome Measure

Many managed care evaluations and statewide initiatives (e.g., Pennsylvania[44]) examine patients within diagnosis-related groups (DRGs).[45] To parallel these approaches, we selected patients hospitalized for medical treatment of a new acute myocardial infarction defined by DRGs. We chose acute myocardial infarction because it is common and has relatively high death rates. All patients had either a principal or secondary five-digit ICD-9 CM discharge diagnosis code beginning with 410 (acute myocardial infarction) and ending with 1 (initial treatment). By including patients with a fifth digit of 1, we felt reasonably comfortable that patients had a new infarct (prior coding guidelines did not distinguish distant acute myocardial infarctions from new events[46]; these rules changed in October 1989). We included patients in DRGs 121 (circulatory disorders with acute myocardial infarction and cardiovascular complication, discharged alive), 122 (circulatory disorders with acute myocardial infarction without cardiovascular complication, discharged alive), and 123 (circulatory disorders with acute myocardial infarction, expired).

Our outcome measure was in-hospital death. The MedisGroups data did not contain information on deaths after discharge.

Analytic Methods

Using each severity measure, we calculated a predicted probability of death for each patient in the sample from a multivariable logistic regression model including the severity score and dummy variables representing a cross-classification of patients by sex and eight age categories (18–44, 45–54, 55–64, 65–69, 70–74, 75–79, 80–84, and 85 years of age or older). Severity scores were entered as

TABLE 3—Number of Times That Pairs of Severity Methods Agreed on Flagging Hospitals as among the 10 Worst

	MG-O	MG-E	PS 1	PS 2	DS-PR	DS-ST	PMCs-SS	CM	APR-DRG	R-DRG	Unadj
MG-O	10	8	8	8	4	7	7	8	6	7	7
MG-E		10	8	10	3	6	5	6	4	5	6
PS 1			10	8	4	6	6	8	5	7	7
PS 2				10	3	6	5	6	4	5	6
DS-PR					10	6	5	4	5	3	4
DS-ST						10	7	5	6	5	6
PMCs-SS							10	6	9	6	6
CM								10	6	8	8
APR-DRG									10	5	6
R-DRG										10	6
Unadj											10

Note. Figures are numbers of hospitals flagged by both methods. Number of hospitals on which pairs of methods agreed and associated κ value: 3, κ = .22; 4, κ = .33; 5, κ = .44; 6, κ = .56; 7, κ = .67; 8, κ = .78; 9, κ = .89; 10, κ = 1.00. MG-O = original version of MedisGroups; MG-E = empirical version of MedisGroups; PS 1 and PS 2 = Physiology Scores 1 and 2; DS-PR = Disease Staging mortality probability; DS-ST = Disease Staging stage; PMCs-SS = Patient Management Categories—Severity Score; CM = Comorbidity Index; APR-DRG = All Patient Refined Diagnosis-Related Groups; R-DRG = refined Diagnosis-Related Groups; Unadj = actual mortality rate, unadjusted for age, sex, or severity of illness.

TABLE 4—Number of Times That Pairs of Severity Methods Agreed on Flagging Hospitals as among the 50 Best

	MG-O	MG-E	PS 1	PS 2	DS-PR	DS-ST	PMCs-SS	CM	APR-DRG	R-DRG	Unadj
MG-O	50	42	44	43	39	41	40	42	39	41	40
MG-E		50	42	45	38	42	39	42	41	39	38
PS 1			50	43	34	39	38	41	39	42	38
PS 2				50	39	41	40	42	40	41	38
DS-PR					50	43	40	38	39	39	40
DS-ST						50	42	42	42	42	43
PMCs-SS							50	42	45	40	39
CM								50	40	42	42
APR-DRG									50	38	38
R-DRG										50	42
Unadj											50

Note. Figures are numbers of hospitals flagged by both methods. Number of hospitals on which pairs of methods agreed and associated κ value: 34, κ = .36; 38, κ = .52; 39, κ = .56; 40, κ = .60; 41, κ = .64; 42, κ = .68; 43, κ = .72; 44, κ = .76; 45, κ = .80. MG-O = original version of MedisGroups; MG-E = empirical version of MedisGroups; PS 1 and PS 2 = Physiology Scores 1 and 2; DS-PR = Disease Staging mortality probability; DS-ST = Disease Staging stage; PMCs-SS = Patient Management Categories—Severity Score; CM = Comorbidity Index; APR-DRG = All Patient Refined Diagnosis-Related Groups; R-DRG = refined Diagnosis-Related Groups; Unadj = actual mortality rate, unadjusted for age, sex, or severity of illness.

either continuous or categorical variables (Table 1). For measures with predicted probabilities of death as scores, we used the logit of the probability as the independent variable in logistic regressions. For severity measures producing continuous scores, additional analyses were performed grouping the continuous scores into 8 to 12 categories entered as dummy variables. We present only findings from the continuous-score models because results were generally similar. To establish a baseline, we report results from a model with only age–sex dummy variables.

Hospital-Level Analyses

For each severity measure, we calculated the expected number of deaths and variance for each of the 100 hospitals. To interpret the observed hospital death rates, we calculated a z score for each hospital as follows: z = (observed death rate − expected death rate)/(standard error of this difference). We ranked hospitals from lowest (fewer deaths than expected) to highest (more deaths than expected) based on these z scores.

We sought to recreate approaches used commonly in report cards[3–7]—identifying statistical outliers, or groups of either particularly problematic or good hospitals. We chose three approaches:

1. Whether the hospital was among the worst 10% (10 highest z scores).
2. Whether the hospital was among the best 50% (50 lowest z scores).
3. Whether the hospital was a statistical outlier (z scores of greater than 2 or less than −2, indicating significantly higher or lower numbers of deaths observed than expected).

For each hospital performance measure, a severity measure either flagged a hospital (e.g., identified it as among the worst 10%) or did not. We counted how often pairs of severity measures agreed about flagging hospitals. For each pair of severity measures, we calculated a kappa statistic based on whether individual hospitals were flagged by one, both, or neither of the two severity measures. Kappa values below 0.4 are interpreted as poor to fair agreement and above 0.7 as excellent.[47] Unadjusted hospital mortality rates were included in these pairwise comparisons.

TABLE 5—Measures of Model Performance for Predicting In-Hospital Death and Percentage of Patients Who Died in the Top Two and Bottom Two Deciles of Predicted Probability of Death

System	*c* (95% CI)	R^2 (95% CI)	% Patients Who Died, by Decile Rank Based on Predicted Probability of Death			
			1	2	9	10
MedisGroups						
Original version	.80 (.80, .81)	.17 (.15, .18)	0.5	1.9	24.4	46.0
Empirical version	.83 (.83, .85)	.23 (.21, .25)	0.5	1.3	25.9	53.7
Physiology Score 1	.82 (.81, .83)	.18 (.16, .20)	0.3	1.4	28.3	46.9
Physiology Score 2	.83 (.82, .84)	.23 (.21, .25)	0.3	1.3	23.8	54.7
Disease Staging						
Mortality probability	.86 (.85, .87)	.27 (.25, .29)	0.3	0.4	26.7	58.4
Stage	.79 (.78, .80)	.17 (.15, .18)	1.4	2.9	22.1	49.7
PMCs—Severity Score	.82 (.81, .83)	.18 (.16, .19)	0.2	0.8	27.3	47.3
Comorbidity Index	.70 (.69, .72)	.06 (.05, .07)	1.4	4.0	24.5	26.3
APR-DRGs	.84 (.83, .85)	.20 (.18, .21)	0.0	0.9	36.3	45.0
R-DRGs	.80 (.78, .81)	.15 (.14, .17)	1.0	1.9	29.6	42.2
Age and sex, interacted	.69 (.68, .70)	.05 (.05, .06)	1.4	4.3	23.2	25.6

Note. APR-DRGs = All Patient Refined Diagnosis-Related Groups; PMCs—Severity Score = Patient Management Categories—Severity Score; R-DRGs = Refined Diagnosis-Related Groups; CI = confidence interval.

Statistical Performance Measures

To understand better the results of the pairwise comparisons, we explored whether severity measures with similar predictive power at the individual patient level agreed better. We used the *c* statistic[48–50] and R^2, commonly reported overall measures of statistical performance.[51] To create 95% confidence intervals around these statistics, we replicated analyses (i.e., fitting the models and calculating performance measures) 80 times using bootstrapping techniques.[52] We also checked for model overfitting using split sample cross-validation.[53] After randomly splitting the data into two subsamples of equal size, we fit each model to each half of the data and computed "validated" statistics (*c* and R^2) based on these models applied to the opposite half. We averaged these two validated statistics to produce each model's cross-validated statistics.

To examine model discrimination, we ranked patients by their predicted probability of death based on each multivariable model. We then divided patients into deciles based on increasing predicted probability of death and report actual death rates among patients in the top and bottom two deciles.

We also calculated two correlation coefficients for each pair of severity measures: the correlation between the predicted probabilities of death at the individual patient level and the correlation between *z* scores at the hospital level.

Results

We studied 11 880 patients from 100 hospitals, with 1574 (13.2%) in-hospital deaths. Patients ranged from 19 to 103 years of age, with a mean age of 68.3 (SD = 13.3) years; 58.1% of patients were male. Length of stay averaged 7.7 (SD = 5.5) days. Ample numbers of diagnosis codes were generally present for rating severity of illness, with a mean of 5.6 (SD = 3.0) diagnosis codes per patient. Only 4.2% of patients had 1 discharge diagnosis; 43.4% had more than 5 diagnosis codes, and 10.2% had 10 or more.

The mean number of patients per hospital was 118.8, with a median of 100 and a range of 33 to 340 patients. The 100 hospitals were generally larger and more involved in teaching than other general acute care institutions nationwide.[41] In the sample, 42 hospitals had more than 300 beds; 14 had less than 100 beds. Only three were public, whereas 96 were private nonprofit; 80 were urban. Thirty-nine had approved residency training programs, and 15 were members of the Council of Teaching Hospitals.

Relative Hospital Performance

Unadjusted mortality rates ranged from 4.8% to 26.4% for the 100 hospitals. After adjusting for age, sex, and severity of illness, 65 facilities had observed mortality rates that were similar to expected rates according to all 10 severity measures. Three hospitals had mortality rates that differed significantly from expected rates according to all 10 severity measures, two with lower rates and one with a higher rate.

For 32 hospitals, observed mortality rates differed significantly from expected rates when judged by 1 or more, but not all 10, severity measures. Sometimes these differences were primarily technical (all 10 *z* scores occupied narrow bands surrounding −2 or 2), but often differences were substantial. Table 2 shows examples of five such hospitals. For instance, 15.4% of hospital B's patients died, ranking it in the 8th decile based on its observed death rate (hospitals in decile 1 had the 10% lowest unadjusted death rates). Three measures found significantly fewer deaths than expected at hospital B, ranking it among the 10% of facilities with the lowest adjusted mortality rates. The other seven severity measures found that hospital B's observed death rate was similar to expected, ranking it from the 1st to 7th deciles by *z* scores.

Tables 3 and 4 show details of comparisons between severity measures on whether hospitals were among the 10% worst or 50% best, indicating the number of hospitals on which agreement occurred; kappa values resulting from

each comparison are shown in footnotes to these tables. The clinical data-based measures tended to agree. The only code-based measure with systematically good agreement with the clinical data-based measures was the Comorbidity Index. The code-based measures varied in their level of agreement with each other. The amount of agreement between the severity measures and the unadjusted model was similar to that between most pairs of severity measures.

On average, individual severity measures identified about 15% of the hospitals as statistical outliers. If 10 measures each independently assign outlier flags to 15 of 100 hospitals, more than 70 would have at least one flag (expected = 80.3; SD < 4). In contrast, if all 10 methods measure the same thing, then 15 identical hospitals would receive 10 outlier flags. As noted, 32 hospitals were flagged as outliers by at least one of the methods, suggesting imperfect but substantial agreement. Kappa analyses showed fair to good agreement across pairs of measures. The average κ associated with comparing outlier status determined by the severity measures with outlier status determined by the unadjusted death rates was 0.53.

Statistical Performance and Correlations

Statistical performance varied across severity systems (Table 5). The *c* statistic and R^2 values had tight confidence intervals. Cross-validated *c* and R^2 values were never more than 0.01 smaller than fitted values. Most models identified groups of patients with very low and very high death rates.

No consistent relationship appeared between agreement among pairs of severity measures on hospital rankings (Tables 3 and 4) and the summary statistical performance measures (Table 5). All Patient Refined Diagnosis-Related Groups (APR-DRGs) and Disease Staging's probability models, which had the highest *c* statistics, demonstrated low agreement in identifying the 10% worst hospitals, agreeing on only five ($\kappa = 0.44$). Despite relatively low statistical performance, the Comorbidity Index and Refined Diagnosis-Related Groups (R-DRGs) had much higher agreement, identifying eight identical hospitals ($\kappa = 0.78$).

Three quarters of the correlation coefficients between pairs of severity measures for predictions of probability of death at the individual patient level were below 0.60; approximately 85% of these correlations were less than 0.70. Correlations between 0.70 and 0.90 were observed between the two versions of MedisGroups, between the two physiology scores, and between the MedisGroups scores and the physiology scores. Other pairs of severity measures generally produced much lower correlations; for example, the correlation between empirical MedisGroups and the R-DRGs was 0.42. In contrast, at the hospital level, differences in predictions for individual patients tended to average out. Therefore, correlation coefficients at the hospital level were higher: about three quarters were above 0.80, and one quarter were above 0.90. At the hospital level, higher correlations occurred among those pairs of severity measures with higher correlations at the individual patient level.

Discussion

Almost one third of the study hospitals were viewed as having significantly better or worse death rates than expected with 1 or more, but not all 10, severity measures. Many pairs of severity measures displayed only fair to good agreement on hospital performance. Thus, whether individual hospitals were identified as especially good or bad frequently depended on the particular severity measure. Evaluations of hospital performance based on in-hospital death rates for acute myocardial infarction are therefore sensitive to the severity-adjustment method.

Six severity measures relied on discharge abstract data, and several had high *c* and R^2 values: Disease Staging's mortality probability measure and the APR-DRGs had higher *c* statistics than either version of MedisGroups and the physiology scores. Given the respective data sources on which these measures are based, these results were not surprising. Although the clinical data-based methods used clinical findings only from the first 2 hospital days, the discharge data-based measures reviewed *all* discharge diagnoses regardless of when they occurred (whether the condition was present on admission or arose later during the hospitalization). Therefore, the ability of code-based measures to predict in-hospital deaths could result from their consideration of catastrophic events, such as cardiac arrest or cardiogenic shock, late in the hospital stay. Preliminary analyses supported this conjecture.[42] For example, 6.0% of patients had the ICD-9 CM cardiac arrest code; 60.4% of them died, compared with 10.2% deaths among patients without cardiac arrest coded. Among patients viewed as sicker by Disease Staging's mortality probability measure than by the empirical MedisGroups measure, 16.2% had cardiac arrest coded, whereas cardiac arrest was coded in only 0.4% of patients seen as sicker by MedisGroups than by Disease Staging.[42] These results suggest that codes such as cardiac arrest are important in discharge abstract-based severity ratings.

Given these findings, use of discharge abstract data to judge quality—as in hospital report cards—raises serious concerns. Discharge diagnoses include all conditions treated during the hospitalization, even events occurring late in the stay possibly due specifically to substandard care.[54–56] To draw conclusions about quality based on severity-adjusted outcomes, it is essential to adjust only for preexisting conditions, not those arising after hospitalization. If, for example, cardiac arrest appears on the discharge abstract, it is impossible to determine whether the patient had cardiac arrest late in the stay due to poor monitoring and care. Hospitals also vary in coding intensity.[57] Despite these major drawbacks of discharge abstract data, many states (e.g., California, Connecticut, Florida, Ohio) and payers used code-based methods to examine hospital mortality rates. Typically, discharge abstracts are the only data readily available across institutions, and they are inexpensive and computer readable. Recently, for example, Iowa switched from MedisGroups to APR-DRGs for their hospital performance reports, largely driven by concerns about MedisGroups' data collection costs.

We included severity measures not originally designed for mortality prediction because they are nonetheless used for this purpose. Hospitals or groups frequently purchase a single severity measure, often at substantial expense, and then use it for multiple activities. Using R-DRGs for mortality prediction is particularly problematic because R-DRGs assign all medical patients dying within 2 days of admission to a low-severity class (they cost less than patients who live). In our study, all of the 48.3% of the 1574 deaths occurring within 2 days were assigned to Refined Diagnosis-Related Group class 0. However, the R-DRGs are used to examine hospital death rates. One local example was a report card produced by the *Boston Globe* to evaluate the quality of Massachusetts hospitals.[40] The *Boston Globe* obtained the state's hospital discharge abstract data set, purchased the

R-DRG software, and then compared individual hospitals' inpatient mortality rates within severity levels defined by R-DRGs to a state average. Because of the R-DRGs' handling of early deaths, the *Boston Globe* dropped all patients who died within 2 days from their analysis—a strategy that clearly affects death rate comparisons across hospitals.

We included the physiology scores not specifically to examine APACHE itself, but because of growing interest in "minimum clinical data sets" containing a small number of well-selected, clinical variables that presumably would be less expensive to collect than a large number of variables. Some states and payers are considering requiring routine reporting of a handful of physiologic values. APACHE weights represent one way to use these physiologic variables. Physiology Scores 1 and 2 used 12 and 17 variables, respectively. MedisGroups' data abstraction protocol considers over 200 clinical findings regardless of patient diagnosis. The physiology scores and the empirical version of MedisGroups performed similarly; for example, the empirical version of MedisGroups and Physiology Score 2 agreed on the 10 worst hospitals. Therefore, at least in cases of acute myocardial infarction, a more parsimonious model (e.g., a physiology score approach) would provide comparable statistical power, provide similar judgments about hospitals, and perhaps cost less to implement.

Judgments about hospital performance based on unadjusted mortality rates agreed nearly as much with assessments by severity measures as did judgments between pairs of severity measures. This result implies that severity adjustment is not useful. Unadjusted rates, however, predict the same 13.2% chance of dying for each patient in every hospital, whereas severity measures identified patients with very different mortality risk (Table 5). Therefore, the severity measures produce information that could help interpret different hospital death rates. For example, reviewers may feel differently about deaths in low-risk vs high-risk patients. In addition, regardless of its statistical impact, severity adjustment is essential to initiating a dialogue with physicians about patient death rates. Otherwise, physicians whose patients die at relatively higher rates will argue, perhaps with good reason, "But my patients are sicker."[5,6]

The results presented here pertain only to acute myocardial infarction patients. We replicated these analyses in two other medical conditions with relatively high in-hospital mortality—pneumonia[36] and stroke[35]—and one surgical condition—coronary artery bypass graft.[41] Although there were important similarities across conditions, there were also differences. For the medical conditions, mortality performance varied depending on which severity measure was used for around 30% of the hospitals, but for coronary artery bypass graft few differences across severity measures appeared important. Especially for coronary artery bypass graft, hospital rankings based on observed death rates were generally unchanged after severity adjustment. The specific severity measures involved in the majority of disagreements about flagging outlier hospitals varied somewhat across conditions; for example, in cases of stroke, Disease Staging's mortality probability method most often disagreed with other severity measures, and in cases of pneumonia, the empirical version of MedisGroups disagreed the most.

Findings relating to statistical performance also varied across conditions. For example, as in cases of acute myocardial infarction, code-based measures had better *c* statistics than the clinical data-based measures for coronary artery bypass graft patients; in cases of coronary artery bypass graft, the best *c* value was 0.85 for R-DRGs followed by 0.83 for APR-DRGs, compared with 0.74 for the empirical version of MedisGroups and 0.73 for the original version of MedisGroups and Physiology Scores 1 and 2.[41] In contrast, somewhat in cases of pneumonia and especially in cases of stroke, the clinical data-based methods outperformed the code-based measures. With pneumonia, the highest *c* value was 0.85 for the empirical version of MedisGroups, followed by 0.83 for R-DRGs and 0.82 for Physiology Score 2.[36] With stroke, the best *c* statistic was 0.87 for the empirical version of MedisGroups followed by 0.84 for Physiology Score 2; the performance of the code-based methods was disappointing (e.g., 0.77 for APR-DRGs and 0.74 for Disease Staging mortality probability and R-DRGs).[35]

This study has important limitations. The 1992 MedisGroups database contains information only from MedisGroups users and hospitals in states requiring MedisGroups. Independent information about data reliability was not available. The clinical information in the data set was specifically gathered for MedisGroups scoring, giving MedisGroups a possible advantage in evaluating statistical performance. The MedisGroups data contained information on only in-hospital deaths. This situation is typical (although the Medicare program[58] and several states keep data on out-of-hospital deaths, this information is rarely available elsewhere). Postdischarge mortality information is useful because it allows one to hold constant the window of observation (e.g., at 30 days after hospital admission). This constancy is critical when comparing mortality across providers with differing discharge practices.[59] For our research question, however, we have no reason to expect that our findings (perceptions about risk-adjusted mortality rates may differ by how severity is measured) would change for 30-day mortality.

Our results suggest that judgments about hospital performance based on severity-adjusted mortality can be sensitive to the severity measure. The 10 severity measures agreed about relative hospital performance more often than if judgments were completely independent. Nonetheless, for an individual hospital, perceptions of mortality performance could vary according to different severity adjustment methods. A comprehensive evaluation of the relative merit of severity measures is beyond the scope of a single article. However, these findings raise important questions for report card efforts to judge hospital performance by using severity-adjusted death rates. Because of the uncertainty surrounding the clinical meaning of severity-adjusted hospital mortality rates, it is important to weigh what actions may reasonably be founded on this information (e.g., decisions about contracting with particular hospitals, designating selected facilities as preferred providers). Using this information to guide punitive actions against providers is not indicated until its meaning is clearer. Even if data appear methodologically sound, publicity surrounding the release of provider-specific findings can have untoward consequences (e.g., hospitals and physicians avoiding high-risk patients).[60–62] In addition, with the exception of a few studies primarily involving Medicare patients[63–66] and some that showed different results across different diseases,[66] there is little evidence that mortality rates are valid measures of hospital quality. □

Acknowledgments

This research was supported by Grant R01 HS06742-03 from the Agency for Health Care Policy and Research. Dr Daley is Senior Research Associate, Career Development Program of the Department of Veterans Affairs Health Services Research and Development Service.

References

1. Epstein A. Performance reports on quality—prototypes, problems, and prospects. *N Engl J Med.* 1995;333:57–61.
2. Kassirer JP. The use and abuse of practice profiles. *N Engl J Med.* 1994;330:634–636.
3. *Health Care Reform. "Report Cards" Are Useful but Significant Issues Need to Be Addressed.* Washington, DC, US General Accounting Office, Health, Education, and Human Services Division; September 1994. GAO/HEHS 94-219.
4. *Employers Urge Hospitals to Battle Costs Using Performance Data.* Washington, DC, US General Accounting Office, Health, Education, and Human Services Division; October 1994. GAO/HEHS 95-1.
5. Iezzoni LI, Shwartz M, Restuccia J. The role of severity information in health policy debates: a survey of state and regional concerns. *Inquiry.* 1991;28:117–128.
6. Iezzoni LI, Greenberg LG. Widespread assessment of risk-adjusted outcomes: lessons from local initiatives. *Joint Committee J Quality Improvement.* 1994;20:305–316.
7. Localio AR, Hamory BH, Sharp TJ, Weaver SL, TenHave TR, Landis JR. Comparing hospital mortality in adult patients with pneumonia. A case study of statistical methods in a managed care program. *Ann Intern Med.* 1995;122:125–132.
8. Iglehart JK. Academic medical centers enter the market: the case of Philadelphia. *N Engl J Med.* 1995;333:1019–1023.
9. Selker HP. Systems for comparing actual and predicted mortality rates: characteristics to promote cooperation in improving hospital care. *Ann Intern Med.* 1993;118:820–822.
10. Kassirer JP. The use and abuse of practice profiles. *N Engl J Med.* 1994;330:634–636.
11. McMahon LF, Billi JE. Measurement of severity of illness and the Medicare prospective payment system: state of the art and future directions. *J Gen Intern Med.* 1988;3:482–490.
12. The Quality Measurement and Management Project. *The Hospital Administrator's Guide to Severity Measurement Systems.* Chicago, Ill: The Hospital Research and Educational Trust of the American Hospital Association; 1989.
13. Iezzoni LI, ed. *Risk Adjustment for Measuring Health Care Outcomes.* Ann Arbor, Mich: Health Administration Press; 1994.
14. *All Patient Refined Diagnosis Related Groups. Definition Manual.* Wallingford, Conn: 3M Health Information Systems; 1993.
15. Edwards N, Honemann D, Burley D, Navarro M. Refinement of Medicare diagnosis-related groups to incorporate a measure of severity. *Health Care Financing Rev.* 1994;16:45–64.
16. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis.* 1987;40:373–383.
17. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol.* 1992;45:613–619.
18. Gonnella JS, Hornbrook MC, Louis DZ. Staging of disease: a case-mix measurement. *JAMA.* 1984;251:637–644.
19. Markson LE, Nash DB, Louis DZ, Gonnella JS. Clinical outcomes management and disease staging. *Eval Health Professions.* 1991;14:201–227.
20. Naessens JM, Leibson CL, Krishan I, Ballard DJ. Contribution of a measure of disease complexity (COMPLEX) to prediction of outcome and charges among hospitalized patients. *Mayo Clin Proc.* 1992;67:1140–1149.
21. Brewster AC, Karlin BG, Hyde LA, Jacobs CM, Bradbury RC, Chae YM. MEDISGRPS: a clinically based approach to classifying hospital patients at admission. *Inquiry.* 1985;12:377–387.
22. Iezzoni LI, Moskowitz MA. A clinical assessment of MedisGroups. *JAMA.* 260:3159–3163.
23. Blumberg MS. Biased estimates of expected acute myocardial infarction mortality using MedisGroups admission severity groups. *JAMA.* 1991;265:2965–2970.
24. Steen PM, Brewster AC, Bradbury RC, Estabrook E, Young JA. Predicted probabilities of hospital death as a measure of admission severity of illness. *Inquiry.* 1993;30:128–141.
25. Young WW, Kohler S, Kowalski J. PMC patient severity scale: derivation and validation. *Health Serv Res.* 1994;29:367–390.
26. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med.* 1985;13:818–829.
27. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. An evaluation of outcome from intensive care in major medical centers. *Ann Intern Med.* 1986;104:410–418.
28. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. *Chest.* 1991;100:1619–1636.
29. Knaus WA, Wagner DP, Zimmerman JE, Draper EA. Variations in mortality and length of stay in intensive care units. *Ann Intern Med.* 1993;118:753–761.
30. Freeman JL, Fetter RB, Park H, et al. Refinement. In: Fetter RB, Brand DA, Gamache D, eds. *DRGs: Their Design and Development.* Ann Arbor, Mich: Health Administration Press; 1991:57–79.
31. Freeman JL, Fetter RB, Park H, et al. Diagnosis-related group refinement with diagnosis- and procedure-specific comorbidities and complications. *Med Care.* 1995;33:806–827.
32. Iezzoni LI, Restuccia JD, Shwartz M, et al. The utility of severity of illness information in assessing the quality of hospital care. *Med Care.* 1992;30:428–444.
33. Thomas JW, Ashcraft MLF. Measuring severity of illness: six severity systems and their ability to explain cost variations. *Inquiry.* 1991;28:39–55.
34. MacKenzie TA, Willan AR, Lichter J, et al. *Patient Classification Systems: An Evaluation of the State of the Art.* Kingston, Ontario, Canada: Case Mix Research, Queens College; 1991.
35. Iezzoni LI, Shwartz M, Ash AS, Hughes JS, Daley J, Mackiernan YD. Using severity-adjusted stroke mortality rates to judge hospitals. *Int J Qual Health Care.* 1995;7:81–94.
36. Iezzoni LI, Shwartz M, Ash AS, Hughes JS, Daley J, Mackiernan YD. Severity measurement methods and judging hospital death rates. *Med Care.* 1996;34:11–28.
37. *Uniform Hospital Discharge Data Minimum Data Set.* Hyattsville, Md: US Dept of Health, Education and Welfare, National Committee on Vital and Health Statistics; 1980. DHEW publication PHS 80-1157.
38. Anderson G, Steinberg EP, Whittle J, Powe NR, Antebi S, Herbert R. Development of clinical and economic prognoses from Medicare claims data. *JAMA.* 1990;263:967–972.
39. Connell FA, Diehr P, Hart LG. The use of large data bases in health care studies. *Annu Rev Public Health.* 1987;8:51–74.
40. Kong D. High hospital death rates. *Boston Globe.* October 3, 1994:1, 6, 7.
41. Iezzoni LI, Shwartz M, Ash AS, et al. *Evaluating Severity Adjustors for Patient Outcome Studies. Final Report.* Boston, Mass: Beth Israel Hospital; May 10, 1995. Report from Agency for Health Care Policy and Research grant R01-HSO6742.
42. Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes. *Ann Intern Med.* 1995;123:763–770.
43. Iezzoni LI, Hotchkin EK, Ash AS, Shwartz M, Mackiernan Y. MedisGroups® databases: the impact of data collection guidelines on predicting in-hospital mortality. *Med Care.* 1993;31:277–283.
44. *Coronary Artery Bypass Graft Surgery: Technical Report, Vol II, 1991.* Harrisburg, Pa: Pennsylvania Health Care Cost Containment Council; February 1994.
45. Vladeck BC. Medicare hospital payment by diagnosis-related groups. *Ann Intern Med.* 1984;100:576–591.
46. Iezzoni LI, Burnside S, Sickles L, Moskowitz MA, Sawitz E, Levine PA. Coding of acute myocardial infarction: clinical and policy implications. *Ann Intern Med.* 1988;109:745–751.
47. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
48. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med.* 1984;3:143–152.
49. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143:29–36.
50. Hanley JA, McNeil BJ. A method of comparing the area under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983;148:839–843.
51. Ash AS, Shwartz M. Evaluating the performance of risk-adjustment methods: dichotomous measures. In: Iezzoni LI, ed. *Risk Adjustment for Measuring Health Care Outcomes.* Ann Arbor, Mich: Health Administration Press, 1994:313–346.
52. Efron B. Bootstrap methods: another look at the jacknife. *Ann Stat.* 1979;7:1–26.
53. Daley J. Validity of risk-adjustment methods. In: Iezzoni LI, ed. *Risk Adjustment for Measuring Health Care Outcomes.* Ann Arbor, Mich: Health Administration Press; 1994:239–262.
54. Iezzoni LI, Foley SM, Heeren T, et al. A method for screening the quality of hospital care using administrative data: prelimi-

nary validation results. *Qual Rev Bull.* 1992;18:361–371.
55. Iezzoni LI, Daley J, Heeren T, et al. Identifying complications of care using administrative data. *Med Care.* 1994;32:700–715.
56. Shapiro MF, Park RE, Keesey J, Brook RH. The effect of alternative case-mix adjustments on mortality differences between municipal and voluntary hospitals in New York City. *Health Serv Res.* 1994;29:95–112.
57. Iezzoni LI, Foley SM, Daley J, Hughes J, Fisher ES, Heeren T. Comorbidities, complications, and coding bias. Does the number of diagnosis codes matter in predicting in-hospital mortality? *JAMA.* 1992;267:2197–2203.
58. Sullivan LW, Wilensky GR. *Medicare Hospital Mortality Information: 1987, 1988, 1989.* Washington, DC: US Dept of Health and Human Services, Health Care Financing Administration; 1991. HCFA publication 00720.
59. Jencks SF, Williams DK, Kay TL. Assessing hospital-associated deaths from discharge data: the role of length of stay and comorbidities. *JAMA.* 1988;260:2240–2246.
60. Byer MJ. Faint hearts. *New York Times.* March 21, 1992:23.
61. Oberman L. Valuable input? Risk-adjusted data credited for better outcomes. *American Medical News.* December 28, 1992.
62. Green J, Wintfeld N. Report cards on cardiac surgeons. Assessing New York state's approach. *N Engl J Med.* 1995;332:1229–1232.
63. Keeler EB, Rubenstein LV, Kahn KL, et al. Hospital characteristics and quality of care. *JAMA.* 1992;268:1709–1714.
64. Kuhn EM, Hartz AJ, Gottlieb MS, Rimm AA. The relationship of hospital characteristics and the results of peer review in six large states. *Med Care.* 1991;29:1028–1038.
65. Hartz AJ, Gottlieb MS, Kuhn EM, Rimm AA. The relationship between adjusted hospital mortality and the results of peer review. *Health Serv Res.* 1993;27:765–777.
66. Thomas JW, Holloway JJ, Guire KE. Validating risk-adjusted mortality as an indicator of quality of care. *Inquiry.* 1993;30:6–22.

Internet Web Site Established for Multidisciplinary Health Care Information

GeoHealthWeb, an international multidisciplinary web site offering a full range of health care information and networking for health care consumers, professionals, educators, and business people, has been established on the Internet. A nonprofit service of the National Center for Computer Education and Research in Healthcare (NCCERH) at the St. Louis College of Pharmacy, GeoHealthWeb provides the first free communication marketplace where virtually any kind of health care information is available with only a computer and modem.

GeoHealthWeb is designed to serve the educational, professional, and informational needs of consumers, physicians, pharmacists, nurses, and allied health care professionals on a global scale. Users pay no fee to register for this service, yet are able to take advantage of a broad range of information and services. Included among GeoHealthWeb's sponsors and participating organizations are approximately 80% of the national pharmacy and health care associations in Canada and the United States, and many health care providers and related businesses.

GeoHealthWeb can be previewed now. As a World Wide Web application on the Internet, the universal resource locator (URL) address for GeoHealthWeb is ≪http://geohealthweb.com≫.

For further information about GeoHealthWeb, contact Renato Cataldo, Jr., PharmD, Director of Microcomputer Applications and Associate Professor of Pharmacy Administration, NCCERH, St. Louis College of Pharmacy, 4588 Parkview Place, St. Louis, MO 63110; e-mail renato@slcop.stlcop.edu.