Pergamon

International Journal for Quality in Health Care, Vol. 7, No. 2, pp. 81–94, 1995

1353–4505/95 $9.50 + 0.00

1353-4505(95)00007-0

Using Severity-adjusted Stroke Mortality Rates to Judge Hospitals

LISA I. IEZZONI,* MICHAEL SHWARTZ,‡ ARLENE S. ASH,§ JOHN S. HUGHES,‖ JENNIFER DALEY*¶ and YEVGENIA D. MACKIERNAN*

*Division of General Medicine and Primary Care, Department of Medicine, Harvard Medical School, Beth Israel Hospital, The Charles A. Dana Research Institute, and The Harvard–Thorndike Laboratory, Boston, MA, USA

‡Health Care Management Program and Operations Management Department, School of Management, Boston University, Boston, MA, USA

§Health Care Research Unit, Section of General Internal Medicine, Evans Memorial Department of Clinical Research and Medicine, Boston University Medical Center, Boston, MA, USA

‖Department of Medicine, West Haven Veterans Affairs Medical Center, West Haven, CT, USA

¶Health Services Research and Development, Department of Medicine, Brockton/West Roxbury Veterans Affairs Medical Center, West Roxbury, MA, USA

Mortality rates are commonly used to judge hospital performance. In comparing death rates across hospitals, it is important to control for differences in patient severity. Various severity tools are now actively marketed in the United States. This study asked whether one would identify different hospitals as having higher- or lower-than-expected death rates using different severity measures. We applied 11 widely-used severity measures to the same database containing 9407 medically-treated stroke patients from 94 hospitals, with 916 (9.7%) in-hospital deaths. Unadjusted hospital mortality rates ranged from 0 to 24.4%. For 27 hospitals, observed mortality rates differed significantly from expected rates when judged by one or more, but not all 11, severity methods. The agreement between pairs of severity methods for identifying the worst 10% or best 50% of hospitals was fair to good. Efforts to evaluate hospital performance based on severity-adjusted, in-hospital death rates for stroke patients are likely to be sensitive to how severity is measured.

Key words: Mortality rates, severity, hospital quality.

INTRODUCTION

Examining mortality rates have become a staple of monitoring hospital performance despite the dearth of firm scientific evidence linking mortality to hospital quality. The literature addressing this relationship provides inconsistent results. Several reports link higher-than-expected mortality rates to substandard care [1–3], while some do not [4,5], and others provide equivocal conclusions [6–9]. Hospital representatives and even some policy-makers question the real meaning of information on hospital death rates [10,11]. Nevertheless, examining mortality rates is likely to remain a centerpiece of assessing hospital performance. One compelling reason is that data on deaths are routinely available, while information on other outcomes (e.g. patients' functional status, quality of life, satisfaction with care) is not. Another reason is that death is easily understood and defined, and it has obvious importance.

Submitted 9 November 1994; accepted 8 February 1995.
Correspondence: L. I. Iezzoni, Division of General Medicine and Primary Care, Department of Medicine, Beth Israel Hospital, 330 Brookline Avenue, Boston, MA 02215, USA.

Despite debates about the merits of death rates, most efforts to compare rates across hospitals realize the importance of controlling for patient risk or severity of disease [12,13]. Severity adjustment recognizes that some hospitals treat sicker patients (e.g. patients at higher risk of imminent death) than others. While few argue with this premise, how best to measure severity is unclear, primarily because of limitations in available data. A variety of severity measurement tools are now available [14–16]. In the United States and other countries, proprietary severity measures are actively marketed to hospitals, health information companies, government representatives and even purchasers of care (e.g. business leaders). Despite the potential impact of these methods on health care providers and patients, relatively little information is available to guide choice of a severity method. Many articles, primarily by the developers, describe individual severity measures and their statistical performance [17–33], but few studies involving multiple severity methods have been reported by independent researchers [34–36]. Given the variety of severity measures, an important question is whether different severity methods would produce different assessments of hospital mortality performance.

This study applied existing methods in the way they are often currently used in the United States and asked the following specific question: would one identify different hospitals as having lower or higher death rates using different severity measures? We applied 11 common severity measures to the same data and focused on in-hospital deaths for patients managed medically for stroke.

METHODS

Severity methods

We considered 11 methods (Table 1) that are representative of approaches used currently in the United States for severity-adjusting outcomes data for state or regional comparisons across hospitals [37–41], for individual hospital activities (e.g. internal quality monitoring, negotiating contracts with managed care organizations), and clinical and health services research into patient outcomes.

Each of the 11 methods has its own definition of severity (Table 1), reflecting how it was derived and calibrated. For example, APR-DRGs were developed to explain resource consumption during hospital stays; the empirical version of MedisGroups was calibrated to predict in-hospital mortality for persons within each of 64 disease groups. Systems assign either numerical scores or values on a continuous scale (Table 1). Seven approaches (all except the two MedisGroups and two Physiology Scores) calculate scores using standard data elements from hospital discharge abstracts [42–44] such as age, sex, and diagnoses and procedures coded using the *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM). MedisGroups and the Physiology Scores assess severity using clinical data abstracted from medical records.

Database

To assign severity scores, computerized algorithms were applied to a data file extracted from the 1992 MedisGroups® Comparative Database. Briefly, the MedisGroups Comparative Database contains the clinical information collected on hospitalized patients during medical record reviews using the MedisGroups severity measure [23–26]. Hospital purchasers of MedisGroups provide these data to its vendor, MediQual Systems, Inc.; they are then merged into a single database representing over a million discharges and approximately 450 hospitals. The 1992 MedisGroups Comparative Database is a subset of this larger file, containing all 743,964 calendar year 1991 discharges from 108 acute-care hospitals. These 108 facilities were chosen by MediQual Systems for their longitudinal database independently of the research reported here, selecting facilities they viewed as having good quality data and representing a range of hospital characteristics.

To ensure an adequate sample size for hospital-level analyses, we eliminated 14 institutions with fewer than 30 stroke cases (a total of 222 patients; range 0–28 patients per hospital). This resulted in a hospital sample of 94 facilities (Table 2). Information on hospital characteristics, both for the sample and for hospitals nationwide, was taken from the American Hospital Association annual survey.

The original and empirical admission Medis-

TABLE 1. Description of eleven severity methods*

System	Source/Vendor	Data used and definition of severity	Classification approach
Discharge abstract-based methods			
Methods with a clinical definition of severity			
Disease Staging [20–22]	SysteMetrics/MEDSTAT Group, Santa Barbara, CA	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 [31]	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.* [18], coded version patterned after Deyo *et al.* [19]	Discharge abstract; risk of death within one year of medical hospitalization	Integer from additive scale representing number and severity of comorbidities
Methods with a resource-based definition of severity			
Acuity Index Method (AIM) [16]	Iaméter, San Mateo, CA	Discharge abstract; length of hospital stay	Scores 1, 2, 3, 4 or 5 within DRG
All Patient Refined Diagnosis Related Groups (APR-DRGs) [17]	3M Health Information Systems, Wallingford, CT	Discharge abstract; total hospital charges	Four severity classes (A, B, C, D) within adjacent DRGs
Refined Diagnosis Related Groups (R-DRGs) [32,33]	Yale University refinement of DRGs provided by Yale Project Director Karen Schneider, Health Systems Consultants, New Haven, CT	Discharge abstract; length of hospital stay, total hospital charges	Three severity classes (B, C, D) within adjacent medical DRGs‡; "early" deaths grouped in lowest severity class
Clinical data-based methods			
MedisGroups (Atlas MQ)	MediQual Systems, Inc., Westborough, MA	Clinical variables	
Original version [23–25]		Clinical instability indicated by in-hospital death	Admission score 0, 1, 2, 3, or 4
Empirical version [26]		In-hospital death	Probability ranging from 0 to 1
Physiology Score 1	Patterned after Acute Physiology Score (APS) of Acute Physiology and Chronic Health Evaluation (APACHE), version II [27,28]	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 [29,30]	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

*Citations in table relate to references listed at the end of manuscript.
‡DRG = diagnosis related group; "adjacent DRGs" are formed by grouping individual DRGs previously split by complications and comorbidities.

TABLE 2. Selected characteristics of hospital sample and hospitals nationwide in the United States

Selected hospital characteristics	Sample ($n = 94$)	Nation ($n = 5344$)
	(percent of hospitals)	
Selected geographic locations		
Middle Atlantic region	60.6	10.0
Pennsylvania	58.5	4.0
Southern region	12.8	37.5
Pacific region	2.1	12.1
Urban location	79.8	55.6
Rural location	20.2	44.4
Bed size		
Less than 100	12.8	43.3
100–300	44.6	38.3
More than 300	42.6	18.4
Ownership*		
Public, government, nonfederal	4.3	25.9
Private, non-profit	94.6	56.1
Private, for profit	1.1	12.2
Teaching status		
Approved residency training program	41.5	18.5
Council of Teaching Hospitals member‡	16.0	6.6

*The hospital sample did not include any federal institutions (e.g. military, veterans facilities).

‡Members of the Council of Teaching Hospitals are generally tertiary care facilities associated with medical schools.

Groups scores were provided by MediQual Systems. Scores for the other methods had to be assigned. The MedisGroups Comparative Database contains standard discharge abstract information listed by hospitals submitting MedisGroups data, including up to 20 ICD-9-CM discharge diagnosis codes and 50 ICD-9-CM procedure codes. It also includes values of key clinical findings (KCFs) from the admission period (generally the first two hospital days) abstracted from medical records during MedisGroups review [23–26]. KCFs generally indicate acute physiologic derangements or clinical abnormalities. We used this KCF information to create Physiology Scores patterned after APACHE II and III. To do so, available physiologic findings were given weights specified by APACHE II and III based on their values (e.g. a pulse of 145 beats/min had a weight of 13 points for APACHE III [28]. As with APACHE, these weights were summed to produce the score. We could not replicate the actual Acute Physiology Scores of APACHE II or III using the MedisGroups data because complete values for the required 12 and 17 physiologic variables, respectively, were not available. MedisGroups truncates data collection in broadly-defined normal ranges, but a previous study demonstrated that a KCF-based physiology score nevertheless performed well compared with the original APACHE II score [45].

For the seven discharge abstract-based severity measures, we assigned only the code-based version of the Charlson comorbidity index [18], using an approach adapted from ICD-9-CM diagnosis codes listed in Deyo *et al.* [19]. Other severity scoring was performed by the vendors of the different systems (Table 1). Based on specifications provided by these vendors, we prepared computer files containing, in the format requested, the necessary discharge abstract data elements extracted from the MedisGroups Comparative Database. Vendor specifications varied slightly, primarily regarding the number of ICD-9-CM codes per case their software could process. For sampling patients (see below), we used the DRG assigned by MediQual Systems, representing Version 9.0 of the Medicare DRGs.

The vendors returned the data to us after scoring. Scores of the different systems were merged into a single analytic file, with a 100% merge rate.

Study sample and outcome measure

The study sample included patients hospitalized for medical treatment of a stroke. We first selected patients from the 1992 MedisGroups Comparative Database with a principal diagnosis of one of the following ICD-9-CM diagnosis codes (X can be any integer value): 433.X (occlusion and stenosis of precerebral arteries); 434.X (occlusion of cerebral arteries); and 436 (acute but ill-defined cerebrovascular disease). We did not include subarachnoid, subdural, extradural or intracerebral hemorrhages. To ensure that we included patients receiving medical treatment, we selected only patients in medical DRG 14 (specific cerebrovascular disorders except transient ischemic attack).

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

Analytical methods

The major research question was: to what extent did hospital performance, indicated by severity-adjusted death rates, depend on the system used to adjust for severity?

Using each severity method, a predicted probability of death was calculated 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). The severity score was entered as either a continuous or categorical variable as indicated in Table 1. For methods with predicted probabilities of death as scores (Disease Staging's probability model and empirical MedisGroups), we used the logit of the probability as the independent severity variable in the logistic regression. For comparison we also reported a model including only the age–sex dummy variables.

Statistical performance measures. The *c* statistic and R^2 are commonly reported as overall measures of the ability of statistical models to predict patient outcomes. The *c* statistic equals the area under a Receiver Operating Characteristic (ROC) curve [46]; it measures how well the model discriminates between patients who lived and those who died (a *c* value of 0.5 indicates no ability to discriminate, while a value of 1.0 indicates perfect discrimination) [47,48]. For each severity method, we also ranked patients by their predicted probability of death based on the multivariable model. We then divided patients into 10 equal groups (i.e. deciles from 1 to 10) based on increasing predicted probability of death. We report the actual death rates among patients in the top and bottom two deciles for each severity method. These figures give a sense of how well the models separated patients with very high and very low risks of death.

Hospital-level analyses. For each severity method, we calculated the expected number of deaths for each of the 94 hospitals. This was done by summing, across patients within each facility, the predicted probability of death (p_i for the ith patient) from the multivariable logistic regression model for the particular severity method. The variance in the number of deaths was calculated as the sum of $p_i(1 - p_i)$ for patients in the hospital. When considering hospital mortality rates unadjusted for age, sex, or severity, p_i was set equal to 0.097, the overall probability of death in the sample. To interpret the observed hospital death rates, we calculated a *z*-score for each hospital as follows: z = (observed number of deaths − expected number of deaths)/(square root of the variance in the number of deaths). We then ranked hospitals from lowest (fewer deaths than expected) to highest (more deaths than expected) based on these *z*-scores. These hospital ranks were divided into equal deciles, from 1 to 10.

We examined three measures of hospital performance as follows:

1. Whether the hospital ranked among the worst 10% of hospitals (the nine hospitals with the highest *z*-scores);
2. Whether the hospital was among the best 50% of facilities (the 47 hospitals with the lowest *z*-scores); and
3. Whether the hospital was a statistical outlier,

defined as z-scores > 2 or < -2, indicating significantly higher or lower numbers of deaths observed than expected.

For each measure, a severity method either "flagged" a hospital (e.g. identified the hospital as among the worst 10%) or it did not. We looked at the number of times pairs of severity methods agreed about flagging a hospital. In addition, for each pair of severity methods, we calculated a kappa statistic based on whether individual hospitals were flagged by one, both or neither of the two severity methods. Kappa measures the extent to which there is more agreement on flagging hospitals by each of the two severity methods than expected by chance. In general, kappa values below 0.4 indicate poor to fair agreement, values between 0.4 and 0.7 moderate to good agreement, and values greater than 0.7 excellent agreement [49]. Unadjusted hospital mortality rates were added as an eleventh method in these pairwise comparisons.

RESULTS

The final data set had 9407 patients from 94 hospitals, with 916 (9.7%) in-hospital deaths. Patients ranged from 18 to 101 years of age, with a mean of 73.8 (standard deviation = 11.8) years of age; 55.9% of patients were female. Length of stay ranged from 1 to 328 days, with a mean of 10.2 (SD = 9.9) days. Most cases had ample numbers of diagnoses codes for rating severity with the discharge abstract-based methods, with a mean of 6.0 (SD = 2.9) diagnosis codes per patient. Just 2.1% of patients had only one discharge diagnosis code; 50.2% had more than five diagnoses listed, and 10.8% had 10 or more diagnoses. The 94 hospitals were generally larger, more urban and not-for-profit, and more involved in teaching than other general acute care institutions in the United States (Table 2). The mean number of patients per hospital was 100.1, with a median of 96 and a range of 31–260 patients.

Measures of statistical performance

The 11 severity systems varied in their statistical performance (Table 3). According to both c and R^2 empirical MedisGroups had the best predictive performance, followed closely by Physiology Score 2. In general, most of the models identified groups of patients with very low death rates (20% of patients with death rates under 5%) and another group with very high death rates (over 15% in the ninth decile

TABLE 3. Measures of model performance for predicting in-hospital death and percent of patients who died in the top two and bottom two deciles of predicted probability of death

System	c Statistic	R^2	Decile rank based on predicted probability of death			
			1	2	9	10
			(percent of patients who died)			
Disease Staging						
Mortality probability	0.74	0.11	3.1	3.5	16.5	32.8
Stage	0.60	0.01	4.8	6.6	14.8	14.3
PMCs—Severity Score	0.73	0.10	2.9	2.7	11.9	36.6
Comorbidity Index	0.61	0.01	4.4	6.0	15.1	15.0
Acuity Index Method	0.66	0.03	2.8	4.3	16.4	18.2
APR–DRGs	0.77	0.10	1.2	3.1	19.0	33.7
R–DRGs	0.74	0.07	2.7	2.6	22.3	26.5
MedisGroups						
Original version	0.80	0.15	0.6	1.7	19.7	33.7
Empirical version	0.87	0.27	0.3	1.1	21.1	49.6
Physiology Score 1	0.80	0.17	0.9	2.6	17.7	42.4
Physiology Score 2	0.84	0.24	0.9	2.3	18.3	48.2
Age and sex, interacted	0.60	0.01	4.6	6.4	14.7	14.6

Table 4. 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
Number died/total number of cases	9/138	15/175	29/228	24/168	15/100
Death rate (percent)	6.5	8.6	12.7	14.3	15.0
Decile rank by unadjusted death rate‡	3	4	9	9	10
z-score (decile rank by z-score)§					
Disease Staging-PR	−1.72 (1)	−2.52 (1)	0.81 (8)	4.62 (10)	1.50 (9)
Disease Staging-Stage	−0.83 (3)	−0.53 (4)	1.73 (10)	1.89 (10)	1.76 (10)
PMC-Severity Score	−1.10 (2)	−0.63 (3)	2.06 (10)	1.34 (9)	2.80 (10)
Comorbidity Index	−0.79 (3)	−0.57 (4)	1.69 (10)	2.00 (10)	1.68 (9)
AIM	−0.92 (3)	−1.03 (3)	1.88 (10)	1.99 (10)	2.42 (10)
APR-DRGs	−1.42 (2)	−1.63 (1)	2.51 (10)	1.31 (9)	3.05 (10)
R-DRG	−1.25 (2)	−0.59 (3)	2.53 (10)	1.79 (10)	1.89 (10)
MedisGroups-Original	−1.83 (1)	−0.55 (4)	2.00 (10)	2.53 (10)	1.34 (9)
Medisgroups–Empirical	−2.54 (1)	−1.66 (1)	1.33 (9)	2.50 (10)	0.81 (8)
Physiological Score 1	−3.08 (1)	−2.63 (1)	1.00 (8)	1.91 (10)	1.76 (9)
Physiological Score 2	−1.84 (1)	−1.38 (2)	1.92 (10)	2.75 (10)	1.21 (9)

*Disease Staging-PR = Disease Staging mortality probability; PMC-Severity Score = Patient Management Category Severity Score; AIM = Acuity Index Method; APR-DRGs = All Patient Refined DRGs; R-DRG = refined DRGs.

‡Decile of rank of hospital by actual death rate, unadjusted for age, sex, or patient severity. 1 = death rate in the lowest 10%; 10 = death rate in the highest 10%.

§Decile of rank of z-score. 1 = z-score in the lowest 10%; 10 = z-score in the highest 10%.

and often close to, or higher than, 35% in the top decile).

Relative hospital performance

Actual mortality rates for the 94 hospitals ranged from 0 to 24.4% (unadjusted for patient age, sex and severity). Ten hospitals had unadjusted death rates under 5%, while nine facilities had death rates that were 15% or higher. After adjusting for age, sex and severity, 67 facilities had observed mortality rates that did not differ significantly from expected according to all 11 severity methods. No hospitals had mortality rates that differed significantly from expected according to all 11 severity methods.

For 27 hospitals, observed mortality rates differed significantly from expected rates when judged by one or more, but not all 11, severity methods. These differences were often more technical than real (e.g. all 11 z-scores occupied a narrow band containing −2), but many differences were substantial. Examples of five such hospitals are shown in Table 4. For instance, in Hospital B, 8.6% (15/175) of patients died, ranking this facility in the 4th decile based on its observed death rate (where hospitals in decile 1 had the lowest unadjusted death rates). Despite this, two severity methods found that Hospital B had significantly fewer deaths than expected (z-scores < -2), and they ranked Hospital B among the 10% of facilities with the lowest adjusted mortality rates. In contrast, the other nine severity methods found that Hospital B's observed death rate was similar to expected, ranking it from the 1st to 4th deciles according to its z-score.

Tables 5 and 6 show details of comparisons between pairs of methods on whether hospitals were in the worst 10% or best 50%, indicating the number of hospitals on which there was agreement and, in footnotes to the tables, the kappa values resulting from each comparison. The clinical data-based methods tended to agree somewhat better with each other than with code-based approaches, but this agreement was not perfect even among measures that were closely related. For example, the original and empirical MedisGroups methods agreed on only five of nine hospitals considered among the worst 10%. The code-based measure that showed systematically better agreement with

TABLE 5. Flagging hospitals as among the worst 10 percent: number of times pairs of severity methods agreed

	DS-PR	DS-ST	PMC-SS	CM	AIM	APR	R-DRG	MG-O	MG-E	PS 1	PS 2	Unadj
	(number of hospitals flagged by both methods)											
DS-PR	9	3	3	4	4	3	4	5	3	5	3	3
DS-ST		9	7	8	8	6	8	6	5	6	5	8
PMC-SS			9	6	7	7	8	5	3	5	3	6
CM				9	7	5	7	7	6	7	6	7
AIM					9	7	8	7	4	6	5	7
APR						9	7	5	4	4	3	5
R-DRG							9	6	4	6	4	7
MG-O								9	5	7	7	5
MG-E									9	7	6	5
PS 1										9	6	6
PS 2											9	5
Unadj												9

DS-PR = Disease Staging mortality probability; DS-ST = Disease Staging stage; PMC-SS = Patient Management Category Severity Score; CM = Comorbidity Index; AIM = Acuity Index Method; APR = All Patient Refined DRGs; R-DRG = refined DRGs; MG-O = original MedisGroups; MG-E = empirical MedisGroups; PS 1 and 2 = Physiology Scores 1 and 2; Unadj = actual mortality rate, unadjusted for age, sex, or severity.

Number of hospitals on which pairs of methods agreed and associated kappa (κ) value: 3, $\kappa = 0.26$; 4, $\kappa = 0.39$; 5, $\kappa = 0.51$; 6, $\kappa = 0.63$; 7, $\kappa = 0.75$; 8, $\kappa = 0.88$.

TABLE 6. Flagging hospitals as among the best 50 percent: number of times pairs of severity methods agreed

	DS-PR	DS-ST	PMC-SS	CM	AIM	APR	R-DRG	MG-O	MG-E	PS 1	PS 2	Unadj
	(number of hospitals flagged by both methods)											
DS-PR	47	34	33	34	34	35	33	34	34	31	32	34
DS-ST		47	42	47	42	38	44	42	36	36	40	46
PMC-SS			47	42	41	40	42	40	35	36	37	42
CM				47	42	38	44	42	36	36	40	46
AIM					47	42	43	40	35	39	38	41
APR						47	40	39	35	38	36	38
R-DRG							47	40	34	37	39	44
MG-O								47	38	40	41	42
MG-E									47	38	40	36
PS 1										47	39	36
PS 2											47	41
Unadj												47

DS-PR = Disease Staging mortality probability; DS-ST = Disease Staging stage; PMC-SS = Patient Management Category Severity Score; CM = Comorbidity Index; AIM = Acuity Index Method; APR = All Patient Refined DRGs; R-DRG = refined DRGs; MG-O = original MedisGroups; MG-E = empirical MedisGroups; PS 1 and 2 = Physiology Scores 1 and 2; Unadj = actual mortality rate, unadjusted for age, sex, or severity.

Number of hospitals on which pairs of methods agreed and associated kappa (κ) value: 31, $\kappa = 0.32$; 32, $\kappa = 0.36$; 33 $\kappa = 0.40$; 34 $\kappa = 0.45$; 35, $\kappa = 0.49$; 36, $\kappa = 0.53$; 37, $\kappa = 0.57$; 38, $\kappa = 0.62$; 39, $\kappa = 0.66$; 40, $\kappa = 0.70$; 41, $\kappa = 0.74$; 42, $\kappa = 0.79$; 43, $\kappa = 0.83$; 44, $\kappa = 0.87$; 45, $\kappa = 0.91$; 46, $\kappa = 0.96$; 47, $\kappa = 1.00$

the clinical data-based measures was the Comorbidity Index. The code-based methods varied in their level of agreement with each other, although Disease Staging's mortality probability method differed most often from other systems. The amount of agreement for identifying the worst 10% of hospitals between the 11 severity methods and the unadjusted model was similar to that between most severity methods (kappa values ranging from 0.26 to 0.88).

On average, individual severity methods

identified about 8.4% of the 94 hospitals (about eight hospitals) as *statistical* outliers (*z*-scores $2 >$ or < -2). If each severity method had flagged outlier hospitals entirely at random, 58 facilities would have been expected to receive at least one flag. In contrast, if the systems were measuring essentially the same thing, about eight identical hospitals would have been flagged as outliers by each of the methods. As noted, 27 hospitals were flagged as outliers by at least one of the methods, suggesting substantially more agreement than expected if the severity methods were completely independent of each other.

The kappa analyses showed fair to good agreement across pairs of methods in flagging *statistical outliers*. The lowest average kappa values for flagging outliers were associated with comparisons of Disease Staging's mortality probability, with kappas ranging from 0.08 (with the Comorbidity Index) to 0.33 (with the original MedisGroups). At the high end were the kappas associated with pairwise comparisons with the Patient Management Categories, ranging from 0.15 (with Disease Staging's probability model) to 0.68 (with APR-DRGs). The kappas associated with comparing outlier status determined by the 11 severity methods versus unadjusted death rates ranged from 0.09 (with empirical MedisGroups) to 0.78 (with the R-DRGs).

No consistent relationship appeared between agreement among pairs of severity measures on hospital rankings (Tables 5 and 6) and the summary statistical performance measures (Table 3). For example, the Comorbidity Index, a method with poor statistical performance, had generally good agreement with many other methods in flagging the worst 10% of hospitals. Empirical MedisGroups, the method with the best statistical performance, often disagreed with other methods in flagging the worst 10% of facilities.

DISCUSSION

Unadjusted mortality rates for stroke patients varied widely across hospitals. Judgments about whether severity-adjusted death rates differed from expected, however, sometimes varied depending on the method used to adjust for severity. For over one-quarter of the 94 hospitals, different severity measures yielded different results: these hospitals would have been viewed as having significantly better or worse death rates than expected based on one or more, but not all 11, severity methods. Whether or not an individual hospital was identified as either especially good or bad depended on the particular method used for severity adjustment. Therefore, efforts to evaluate hospital performance based on severity-adjusted, in-hospital death rates for stroke patients are likely to be sensitive to how severity it measured. Several points generate further discussion.

Purpose of severity methods

We included several severity methods that were explicitly intended for purposes other than predicting mortality (Table 1). The statistical performance of some of the resource-based measures were relatively good. For example, APR-DRGs had a better *c* statistic than either Disease Staging model. In this study, however, we could not answer the question about which measure better reflected concerns about quality of care.

In the United States, especially in health policy discussions, no single definition of severity is uniformly applied by all participants [50]. Severity must be defined in terms of a specific outcome and different endpoints interest different persons. For example, business leaders and health care managers often think of severity in terms of resource consumption, while health services providers think of death, functional impairment, quality of life, and other clinical endpoints.

The applications and usefulness of a severity method clearly relate to how it defines severity [51]. However, in the sometimes frenetic rush to quantify hospital performance, these distinctions are frequently forgotten. Hospitals or groups purchase a single severity system, often at substantial expense, and then use it for a variety of purposes, including ones for which it was not designed. A local example involved a recent newspaper article by the *Boston Globe* Spotlight Team presenting its own "Report Card" on Massachusetts hospitals [52]. A *Boston Globe* reporter obtained a public hospital discharge abstract data set, purchased the R-

DRG software, and then examined individual hospitals' inpatient mortality rates within severity levels defined by R-DRGs. The *Boston Globe* published names of 10 hospitals flagged as having poor mortality performance.

Use of the R-DRGs for mortality prediction in particularly problematic because R-DRGs assign all medical patients who die within 2 days of admission to a low severity class (because they cost less than patients who live). Therefore, in our study, all 121 of the 916 deaths that occurred within 2 days were all assigned to R-DRG class 0. In the *Boston Globe* study, the reporter simply dropped all persons who died within 2 days from her analysis [52]—a strategy that has worrisome implications for comparisons of death rates across hospitals.

Concerns about hospital discharge abstract data

Seven methods employed in our study relied upon discharge abstract data, primarily discharge diagnosis codes. Using discharge abstract data to make inferences about hospital quality raises significant concerns, largely because of questions about the timing of conditions that could represent iatrogenic events. To draw conclusions about quality based on severity-adjusted outcomes, it is essential to adjust only for pre-existing conditions, not those arising after hospitalization [25]. Discharge diagnoses include all conditions treated during the hospital stay, regardless of whether they were present on admission or occurred subsequently, possibly due to substandard care [53]. For example, if coma or cardiac arrest appears on the discharge abstract, it is impossible to determine whether it occurred at admission or later in the stay, perhaps due to poor care. One study found that different hospitals would be judged as having worse mortality performance depending on which discharge diagnoses were used to adjust for risk (all diagnoses versus those representing conditions unlikely to arise newly as a result of quality shortfalls) [54].

Another concern pertains to variability in ICD-9-CM coding practices across hospitals. For example, although cardiac arrest does not invariably result in death, the heart always stops when patients die. If some hospitals code cardiac arrest for all deaths and if the severity measure controls for cardiac arrest, this will inflate the expected death rates for such facilities. Comparisons of observed to expected death rates for these institutions are thus likely to appear favourable, due in large measure to this coding practice.

Despite these major problems of discharge abstract data, many States in the United States and health care insurers are nonetheless using code-based severity methods to examine hospital mortality rates [37–41]. In most instances, discharge abstracts are the only data available. They are also computerized and thus easy to score, while abstracted clinical data are more expensive and cumbersome to acquire. Obviously, conclusions about hospital quality based on such data must recognize the possibility of serious inaccuracies due to data limitations. Nevertheless, the perceived urgent need for hospital performance data often leads to concessions about the quality of the underlying data.

Comparing death rates

Judgments of hospital performance based on unadjusted mortality rates agreed nearly as well with assessments of severity methods (e.g., about which hospitals were the worst 10%) as did judgments between pairs of the 11 severity methods. On one hand, this result suggests that, given the current state of the art, severity-adjustment is not useful. However, it is important to note that unadjusted rates predict that each patient within a hospital has the same chance of dying (0.097). As indicated in Table 3, the 11 severity methods are clearly able to identify categories of patients with very different death rates. Therefore, the severity methods produce information that could be valuable for targeting evaluations of care and hospitals with different death rates. For example, one might feel differently if a hospital's deaths are occurring among patients with low predicted probabilities of dying than among those with higher predicted probabilities.

One strategy for improving the credibility of risk-adjusted mortality data as a measure of hospital quality is to look at hospital performance over time (e.g. if a hospital has higher-than-expected death rates year after year, this

creates a stronger suspicion of problems). This was the rationale of the last United States federal publication of hospital death rates based on Medicare administrative data, which included information from three consecutive years [55]. However, in such longitudinal comparisons, it is important to hold constant the severity methodology so that year-to-year differences cannot be attributed to methodologic changes. Most of 11 severity measures are revised periodically by their developers, and new versions may produce slightly different perceptions than the original systems. for example, we included both the original and new empirical versions of MedisGroups, the severity method required for all Pennsylvania hospitals since 1986 [37–41]. The original versions of MedisGroups were diagnosis-independent and derived primarily using clinical judgment [23,24], the revised version, released in 1993, is diagnosis-specific and derived empirically using logistic regression techniques [26]. The kappa associated with agreement in flagging hospitals as *statistical outliers*, using two MedisGroups methods, was only 0.32 (indicating fair agreement). These differences must be considered in longitudinal analyses using different versions of the same severity method.

Limitations of the study

This study has important limitations, especially pertaining to the data set. The 1992 MedisGroups Comparative Database contains information only from self-selected purchasers of MedisGroups or from hospitals in States with MedisGroups data collection mandates. Not surprisingly, therefore, 58.5% of the hospitals were from Pennsylvania. Independent information about data reliability or differences among hospitals in data quality was not available: MediQual Systems indicates that only hospitals with high data quality are included in the Comparative Database. The clinical information in the data set was specifically gathered for MedisGroups scoring, not for use by other severity methods. This may give MedisGroups an advantage in comparisons of statistical performance. Information on routine physiologic findings was not recorded outside MedisGroups-defined normal ranges [45]; certain clinical variables required for other severity methods were not collected. Because of this, we could not examine other clinical data-based severity scores, such as the Computerized Severity Index [56]. Nevertheless, the MedisGroups Comparative Database is unique, including reasonably detailed clinical data on numerous cases, regardless of insurer, from a range of hospitals across the United States. The most important discharge abstract data elements (ICD-9-CM diagnosis codes) are relatively extensive: the 20 diagnosis and 50 procedure coding slots far outstrip the nine diagnosis and six procedure slots available on Medicare billing records.

We included the physiology scores not specifically to examine APACHE itself, but because of the increasing interest in the United States in creating "minimum clinical data sets" containing a small number of well-selected, physiologic variables. For example, States might require that temperature, blood pressure, heart rate, hematocrit and a handful of other values be added to the routinely-reported discharge abstract. APACHE weights represent one way to use these minimum physiologic variables, but there are certainly other approaches. Physiology Score 1 considered 12 variables, while Physiology Score 2 involved 17 items. It is important to note that the MedisGroups data abstraction protocol collects a generic set of over 250 potential "key clinical findings" regardless of patient diagnosis. As shown in Tables 3 through 6, the performance of the Physiology Scores was often comparable to that of empirical MedisGroups.

The MedisGroups data contained information only on in-hospital deaths. This situation is typical: although the Medicare program and several States keep data on out-of-hospital deaths, this information is rarely available elsewhere. Nevertheless, information on post-discharge mortality information is useful because it allows one to hold constant the "window of observation" (e.g. at 30 days following hospital admission). This is critically important if one's purpose is to compare mortality outcomes across providers with differing discharge practices and lengths of stay [57]. However, our primary goal was not to compare outcome rates across providers; it was to explore if different severity adjustment methods yielded different assessments of whether outcomes were better

or worse than expected for specific, individual hospitals. Our main comparison was across the 11 severity systems within individual hospitals (Hospital A judged using Disease Staging versus using empirical MedisGroups; Table 4). We have no reason to expect that our overall finding—that perceptions about severity-adjusted mortality rates may differ by how severity is measured—would change if we had looked instead at 30-day mortality.

Finally, this paper does not provide a comprehensive comparative evaluation of the severity systems. To do so requires attention to a wide range of issues, such as whether there are inherent biases in predictions for certain types of patients and construct validity (the extent to which the risk factors incorporated in the severity methodology include the universe of potential risk factors) [58]. Severity methods may also perform differently in different diseases. In our larger research study, we produced comparable results when examining in-hospital deaths for acute myocardial infarction and pneumonia patients, but we found little difference across severity methods in an analysis of hospitals' coronary artery bypass surgery death rates.

CONCLUSIONS

Our results suggest that judgments about hospital performance based on severity-adjusted stroke mortality rates could be sensitive to the severity method. The 11 severity methods often agreed about relative hospital performance, but for an individual hospital, judgments about stroke mortality could vary using different methods for severity adjustment. Of particular concern is that none of the severity measures was based on extensive data on chronic functional impairments—a clinical consideration likely to be very important in predicting death from stroke. The MedisGroups KCFs focus on acute motor deficits, and ICD-9-CM certainly offers little insight into functional status. Given this uncertainty surrounding the clinical meaning of severity-adjusted mortality rates, it is unclear what conclusions one can reasonably draw about hospital quality from this information.

Acknowledgements: This research was supported by the Agency for Health Care Policy and Research, under grant no. RO1 HS06742-03. Dr. Daley is Senior Research Associate, Career Development Program of the Department of Veterans Affairs Health Services Research and Development Service. The views expressed are solely those of the authors.

REFERENCES

1. Keeler E B, Rubenstein L V, Kahn K L, Draper D, Harrison E R, McGinty M J, Rogers W H and Brook R H, Hospital characteristics and quality of care. *JAMA* **268**: 1709, 1992.
2. Kuhn E M, Hartz A J, Gottlieb M S and Rimm A A, The relationship of hospital characteristics and the results of peer review in six large states. *Med Care* **29**: 1028, 1991.
3. Hartz A J, Kuhn E M, Green R and Rimm A A, The use of risk-adjusted complication rates to compare hospitals performing coronary artery bypass surgery or angioplasty. *Int J Technol Assessment Health Care* **8**: 524, 1992.
4. Park R E, Brook R H, Kosecoff J, Keesey J, Rubenstein L, Keeler E, Kahn K L, Rogers W H and Chassin M R, Explaining variations in hospital death rates. Randomness, severity of illness, quality of care. *JAMA* **264**: 484, 1990.
5. Best W R and Cowper D C, The ratio of observed-to-expected mortality as a quality of care indicator in non-surgical VA patients. *Med Care* **32**: 390, 1994.
6. Dubois R W, Rogers W H, Moxley J H, Draper D and Brook R H, Hospital inpatient mortality. Is it a predictor of quality? *N Engl J Med* **317**: 1674, 1987.
7. Dubois R W and Brook R H, Preventable deaths: who, how often and why? *Ann Intern Med* **109**: 582, 1988.
8. Fink A, Yano E M and Brook R H, The condition of the literature on differences in hospital mortality. *Med Care* **27**: 315, 1989.
9. Thomas J W, Holloway J J and Guire K E, Validating risk-adjusted mortality as an indicator for quality of care. *Inquiry* **30**: 6, 1993.
10. Berwick D M and Wald D L, Hospital leaders' opinions of the HCFA mortality data. *JAMA* **263**: 247, 1990.
11. Podolsky D and Beddingfield K T, America's best hospitals. *US News and World Rep* July 12, **115**: 66, 1993.
12. Selker H P, Systems for comparing actual and predicted mortality rates: characteristics to promote cooperation in improving hospital care. *Ann Intern Med* **118**: 820, 1993.
13. Kassirer J P, The use and abuse of practice profiles. *N Engl J Med* **330**: 634, 1994.
14. McMahon L F and Billi J E, Measurement of severity of illness and the Medicare prospective payment system: state of the art and future directions. *J Gen Int Med* **3**: 482, 1988.
15. The Quality Measurement and Management Project, *The Hospital Administrator's guide to*

severity measurement systems. The Hospital Research and Educational Trust of the American Hospital Association, Chicago, 1989.
16. Iezzoni L I (editor), *Risk adjustment for measuring health care outcomes*. Health Administration Press, Ann Arbor, MI, 1994.
17. *All Patient Refined Diagnosis Related Groups. Definition Manual* 3M Health Information Systems, Wallingford, CT, 1993.
18. Charlson M E, Pompei P, Ales K L and MacKenzie C R, A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chron Dis* **40**: 373, 1987.
19. Deyo R A, Cherkin D C and Ciol M A, Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* **45**: 613, 1992.
20. Gonnella J S, Hornbrook M C and Louis D Z, Staging of disease: a case-mix measurement. *JAMA* **251**: 637, 1984.
21. Markson L E, Nash D B, Louis D Z and Gonnella J S, Clinical outcomes management and disease staging. *Eval Health Prof* **14**: 201, 1991.
22. Naessens J M, Leibson C L, Krishan I and Ballard D J, Contribution of a measure of disease complexity (COMPLEX) to prediction of outcome and charges among hospitalized patients. *Mayo Clin Proc* **67**: 1140, 1992.
23. Brewster A C, Karlin B G, Hyde L A, Jacobs C M, Bradbury R C and Chae Y M, MEDISGRPS: A clinically based approach to classifying hospital patients at admission. *Inquiry* **12**: 377, 1985.
24. Iezzoni L I and Moskowitz M A, A clinical assessment of MedisGroups *JAMA* **260**: 3159, 1988.
25. Blumberg M S, Biased estimates of expected acute myocardial infarction mortality using MedisGroups admission severity groups. *JAMA* **265**: 2965, 1991.
26. Steen P M, Brewster A C, Bradbury R C, Estabrook E and Young J A, Predicted probabilities of hospital death as a measure of admission severity of illness. *Inquiry* **30**: 128, 1993.
27. Knaus W A, Draper E A, Wagner D P and Zimmerman J E, APACHE II: A severity of disease classification system. *Crit Care Med* **13**: 818, 1985.
28. Knaus W A, Draper E A, Wagner D P and Zimmerman J E, An evaluation of outcome from intensive care in major medical centers. *Ann Intern Med* **104**: 410, 1986.
29. Knaus W A, Wagner D P, Draper E A, Zimmerman J E, Bergner M, Bastos P G, Sirio C A, Murphy D J, Lotring T, Damiano A and Harrell F E, The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* **100**: 1619, 1991.
30. Knaus W A, Wagner D P, Zimmerman J E and Draper E A, Variations in mortality and length of stay in intensive care units. *Ann Intern Med* **118**: 753, 1993.
31. Young W W, Kohler S and Kowalski J, PMC patient severity scale: derivation and validation. *Health Serv Res* **29**: 367, 1994.
32. Freeman J L, Fetter R B, Park H, Schneider K C, Lichtenstein J L, Bauman W A, Duncan C C, Hughes J S, Freeman D H and Palmer G R, Refinement. In *DRGs: their design and development*. Fetter R B, Brand D A and Gamache D (Eds) Health Administration Press, Ann Arbor, MI. 57, 1991.
33. Health Systems Management Group School of Organization and Management, Yale University. *DRG Refinement with Diagnostic Specific Comorbidities and Complications: A Synthesis of Current Approaches to Patient Classification*. New Haven, CT: Yale University; Prepared for the Health Care Financing Administration, under Cooperative Agreement Numbers 15-C-98930/1-01 and 17-C-98930/1-025, 1989.
34. Thomas J W, Ashcraft M L F and Zimmerman J E, *An evaluation of alternative severity of illness measures for use by university hospitals. Volume II. Technical Report*. Department of Health Services Management and Policy, School of Public Health, The University of Michigan, Ann Arbor, MI, 1986.
35. Thomas J W and Ashcraft M L F, Measuring severity of illness: six severity systems and their ability to explain cost variations. *Inquiry* **28**: 39, 1991.
36. MacKenzie T A, Willan A R, Lichter J, Greenaway-Coates A, Comis J, Fielding J, Gerlach J, Mahon A and Doran R, *Patient classification systems: an evaluation of the state of the art*. Case Mix Research, Queens College, Kingston, Ontario, 1991.
37. Iezzoni L I, Shwartz M and Restuccia J, The role of severity information in health policy debates: a survey of state and regional concerns. *Inquiry* **28**: 117, 1991.
38. Iezzoni L I and Greenberg L G, Widespread assessment of risk-adjusted outcomes: lessons from local initiatives. *Joint Comm J Qual Improv* **20**: 305, 1994.
39. Iezzoni L E and Greenberg L G, Risk adjustment and current health policy debates. In Iezzoni L I, ed. *Risk adjustment for measuring health care outcomes*. Health Administration Press, Ann Arbor, MI, 347, 1994.
40. United States General Accounting Office; Health, Education, and Human Services Division, *Employers urge hospitals to battle costs using performance data*. Washington, D.C. (GAO/HEHS-95-1). October 1994.
41. United States General Accounting Office; Health, Education, and Human Services Division, *Health care reform: "Report cards" are useful but significant issues need to be addressed*. Washington, D.C. (GAO/HEHS-94-219). September 1994.
42. U.S. Department of Health, Education and Welfare, National Committee on Vital and Health

Statistics, *Uniform Hospital Discharge Data Minimum Data Set*. Hyattsville, MD: DHWQ Pub. No (PHS) 80-1157, 1980.
43. Anderson G, Steinberg E P, Whittle J, Powe N R, Antebi S and Herbert R, Development of clinical and economic prognoses from Medicare claims data. *JAMA* **263**: 967, 1990.
44. Connell F A, Diehr P and Hart L G, The use of large data bases in health care studies. *Ann Rev Public Health* **8**: 51, 1987.
45. Iezzoni L I, Hotchkin E K, Ash A S, Shwartz M and Mackiernan Y, MedisGroups® databases: the impact of data collection guidelines on predicting in-hospital mortality. *Med Care* **31**: 277, 1993.
46. Harrell F E, Lee K L, Califf R M, Pryor D B and Rosati R A, Regression modelling strategies for improved prognostic prediction. *Stat Med* **3**: 143, 1984.
47. Hanley J A and McNeil B J, The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiol* **143**: 29, 1982.
48. Hanley J A and McNeil B J, A method of comparing the area under receiver operating characteristic curves derived from the same cases. *Radiol* **148**: 839, 1983.
49. Landis J R and Koch G G, The measurement of observer agreement for categorical data. *Biometrics* **33**: 159, 1977.
50. Gertman P M and Lowenstein S, A research paradigm for severity of illness: issues for the Diagnosis-Related Group system. Health Care Fin Rev (annual suppl.) 79, 1984.
51. Hornbrook M C, Techniques for assessing hospital care mix. *Ann Rev Public Health* **6**: 295, 1985.
52. Kong D, High hospital death rates. *Boston Globe*. October 3, 1994: 1, 6, 7.
53. Iezzoni L I, Daley J, Heeren T, Foley S M, Fisher E S, Duncan C, Hughes J S and Coffman G A, Identifying complications of care using administrative data. *Med Care* **32**: 700, 1994.
54. Shapiro M F, Park R E, Keesey J and Brook R H, The effect of alternative case-mix adjustments on mortality differences between municipal and voluntary hospitals in New York City. *Health Serv Res* **29**: 95, 1994.
55. Sullivan L W and Wilensky G R, *Medicare hospital mortality information*, 1987, 1988, 1989, Washington, D.C.: U.S. Department of Health and Human Services, Health Care Financing Administration, 1991.
56. Horn S D, Sharkey P D, Buckle J M, Backofen J E, Averill R F and Horn R A, The relationship between severity of illness and hospital length of stay and mortality. *Med Care* **29**: 305, 1991.
57. Jencks S F, Williams D K and Kay T L, Assessing hospital-associated deaths from discharge data: the role of length of stay and comorbidities. *JAMA* **260**: 2240, 1988.
58. Daley J, Validity of risk-adjustment methods. In Iezzoni L I, (Ed.) *Risk adjustment for measuring health care outcomes*. Health Administration Press, Ann Arbor, MI, 239, 1994.