

## The Risks of Risk Adjustment

RECEIVED

ORIGINAL: 1995 - MIZNER  
 Lisa I. Iezzoni, MD, MSc COPIES: de Bien, Harris, Sandusky,

199 FEB 23 AM 11:55

**Context.**—Risk adjustment is essential before comparing patient outcomes across hospitals. Hospital report cards around the country use different risk adjustment methods.

**Objectives.**—To examine the history and current practices of risk adjusting hospital death rates and consider the implications for using risk-adjusted mortality comparisons to assess quality.

**Data Sources and Study Selection.**—This article examines severity measures used in states and regions to produce comparisons of risk-adjusted hospital death rates. Detailed results are presented from a study comparing current commercial severity measures using a single database. It included adults admitted for acute myocardial infarction (n=11 880), coronary artery bypass graft surgery (n=7765), pneumonia (n=18 016), and stroke (n=9407). Logistic regressions within each condition predicted in-hospital death using severity scores. Odds ratios for in-hospital death were compared across pairs of severity measures. For each hospital, z scores compared actual and expected death rates.

**Results.**—The severity measure called Disease Staging had the highest c statistic (which measures how well a severity measure discriminates between patients who lived and those who died) for acute myocardial infarction, 0.86; the measure called All Patient Refined Diagnosis Related Groups had the highest for coronary artery bypass graft surgery, 0.83; and the measure, MedisGroups, had the highest for pneumonia, 0.85 and stroke, 0.87. Different severity measures predicted different probabilities of death for many patients. Severity measures frequently disagreed about which hospitals had particularly low or high z scores. Agreement in identifying low- and high-mortality hospitals between severity-adjusted and unadjusted death rates was often better than agreement between severity measures.

**Conclusions.**—Severity does not explain differences in death rates across hospitals. Different severity measures frequently produce different impressions about relative hospital performance. Severity-adjusted mortality rates alone are unlikely to isolate quality differences across hospitals.

JAMA. 1997;278:1600-1607

IN 1864, LONDON physicians reacted derisively when William Farr from the Registrar General's office released comparisons of death rates across English hospitals.<sup>1</sup> Not only had Farr used questionable statistical methods, they argued, but he also had failed to account for differences in patient characteristics. As one critic asserted: "Any comparison which ignores the difference between the apple-cheeked farm-laborers who seek relief at Stoke Pogis (probably for rheumatism and sore legs), and the wizened [sic], red-herring-like mechanics of Soho or Southwark, who come from a London Hospital, is fallacious."<sup>2</sup>

Over 130 years later, few would contemplate comparing patient outcomes

across hospitals without minimal adjustment for differences in patients' risks. Observers generally agree that some hospitals treat more higher-risk patients than others and that hospitals should not be penalized for accepting risky patients. In addition, without risk adjustment, hospitals with poor outcomes can argue, "But my patients are sicker."

Despite this general consensus about the need for risk adjustment when comparing outcomes, as in publicly released hospital report cards,<sup>3</sup> the devil is in the details. Determining which risk factors to include and how to measure them generates controversy. Practical constraints present difficulties. While researchers have devised detailed risk adjusters,<sup>4,5</sup> these methods are often too expensive to apply in large-scale report card efforts for hospitals. Finally, what do risk-adjusted outcomes tell us?

This article reviews issues raised by risk adjustment for publicly comparing outcomes across providers. Because most outcomes-based report cards to

date have compared hospital death rates,<sup>6-11</sup> I concentrate on this scenario, but the general principles pertain to risk adjustment in other settings as well.

## WHY RISK ADJUST?

Hospitals vary, sometimes widely, in their death rates. The rationale for risk adjustment is to remove one source of this variation, leaving residual differences to reflect quality. The underlying assumption is that outcomes result from a complex mix of factors: patient outcomes equal effectiveness of treatments plus patient risk factors that affect response to treatment plus quality of care plus random chance.

Controlling for patient risk allows us to begin isolating quality differences. But "risk adjustment" is a meaningless phrase without first answering the question: risk of what? Identical risk factors may have different relationships to different outcomes (ie, an attribute suggesting high risk of one outcome may indicate low risk of another outcome). For example, when refining diagnosis related groups (DRGs) to improve their sensitivity to severity for hospital payment, researchers found that medical patients dying within 2 days of admission had relatively low-cost hospitalizations.<sup>18</sup> A risk adjustor that predicts one outcome (eg, death) may not predict another outcome (eg, costs).

Many diverse patient attributes affect risks, including age, sex, acute physiological stability, principal diagnosis (ie, reason for hospitalization) and its severity, the extent and complexity of comorbid illnesses, functional status, psychosocial and cultural factors, socioeconomic characteristics, and preferences for specific outcomes.<sup>19</sup> Measuring certain attributes is challenging and potentially costly (via requiring patient surveys or extensive medical record reviews). For example, physicians and patients often have widely divergent perceptions of the patient's functioning.<sup>20</sup> Whose perspective should be included? In addition, some factors affect outcomes not because of physiology but because of differences in the way people are treated. Black patients<sup>21</sup> and uninsured patients<sup>22</sup> may receive lower-quality hospital care than others. When using risk-adjusted outcomes to evaluate quality, adjusting for race or payer could mask these important differences.

From the Department of Medicine, Division of General Medicine and Primary Care, the Charles A. Dana Research Institute, and the Harvard-Thorndike Laboratory, Harvard Medical School, and Beth Israel Deaconess Medical Center, Boston, Mass.

Reprints: Lisa I. Iezzoni, MD, MSc, Division of General Medicine and Primary Care, Department of Medicine, Beth Israel Deaconess Medical Center, 330 Brookline Ave, East Campus LY-326, Boston, MA 02215.

Table 1.—Examples of Severity Measures for Comparing Hospital Death Rates\*

Severity Measure	Source or Developer	Data Used and Patient Population	Classification Approach
<b>Clinical Data-Based Methods</b>			
APACHE III <sup>23,24</sup>	APACHE Medical Systems, Inc, McLean, Va	17 Physiological variables. Intensive care unit patients	Integer scores from 0 to 299
CHQC <sup>17</sup>	Academy of Medicine of Cleveland, Cleveland, Ohio, and Michael Pine & Associates, Chicago, Ill	Disease-specific clinical variables collected on patients with acute myocardial infarction (AMI), congestive heart failure, pneumonia or chronic obstructive pulmonary disease, stroke, gastrointestinal hemorrhage, large bowel resection, or coronary artery bypass graft (CABG)	Probability of in-hospital death calculated within disease groups
CSI <sup>25,26</sup>	International Severity Information Systems, Salt Lake City, Utah	Disease-specific clinical variables within about 800 disease groups. All hospitalized patients	Scores 1, 2, 3, or 4 for each individual disease; scores 1, 2, 3, or 4 for all diseases combined; "continuous scores" (integer $\geq 0$ ) for all diseases combined
MMPS <sup>27</sup>	Health Care Financing Administration, Baltimore, Md	Disease-specific clinical variables for patients with AMI, congestive heart failure, pneumonia, or stroke	Probability of death 30 d after admission, calculated within disease groups
MedisGroups <sup>28,29</sup>	MediQual Systems, Inc, Westborough, Mass	Over 250 key clinical findings collected on all hospitalized patients	Probability of in-hospital death, calculated within 67 disease groups
CSRS <sup>31,34</sup>	New York Department of Health, Albany	Preoperative CABG risk factors, CABG patients only	Probability of in-hospital death
NNECVDSG <sup>35,36</sup>	Dartmouth-Hitchcock Medical Center, Hanover, NH	Preoperative CABG risk factors, CABG patients only	Probability of in-hospital death
<b>Discharge Abstract-Based Methods</b>			
AIM	Iaméter, San Mateo, Calif	Discharge abstract; all hospitalized patients	Scores 1, 2, 3, 4, or 5 within diagnosis related groups
APR-DRGs <sup>37,38</sup>	3M Health Information Systems, Wallingford, Conn	Discharge abstract; all hospitalized patients	382 Base diagnosis related groups. All except 2 divided into 4 complexity subclasses (1=minor, 2=moderate, 3=major, 4=extreme); 1528 subclasses
California Hospital Outcomes Project <sup>12,13,16</sup>	Office of Statewide Health Planning and Development, Sacramento, Calif	Discharge abstract; patients admitted with AMI	Probability of in-hospital death
Disease Staging Mortality Probability <sup>39,40</sup>	SysteMetrics/MEDSTAT Group, Ann Arbor, Mich	Discharge abstract; all hospitalized patients	Probability of in-hospital death
PMCs Severity Scale <sup>41</sup>	Pittsburgh Health Research Institute at Duquesne University, Pittsburgh, Pa	Discharge abstract; all hospitalized patients	Score of 1, 2, 3, 4, 5, 6, or 7

\*"Clinical data" indicates clinical information (eg, vital signs, test results) abstracted from the medical record; "discharge abstract," standard hospital discharge data elements, such as basic demographics, diagnosis and procedure codes, dates, admission source, and discharge disposition; APACHE, Acute Physiology and Chronic Health Evaluation; CHQC, Cleveland Health Quality Index; CSI, Computerized Severity Index; MMPS, Medicare Mortality Predictor System; CSRS, Cardiac Surgery Reporting System; NNECVDSG, Northern New England Cardiovascular Disease Study Group; AIM, Acuity Index Method; APR-DRGs, All Patient Refined Diagnosis Related Groups; and PMCs, Patient Management Categories.

## WHERE TO START?

Most report card efforts trade off detailed, clinical risk assessments for logistical feasibility and reasonable cost. In addition, designing a method from scratch is very expensive, as some initiatives (eg, the Cleveland Health Quality Choice program<sup>17</sup>) have learned. Thus, taking a measure "off the shelf" is appealing. Many methods now exist specifically to predict in-hospital mortality for comparing hospital outcomes (Table 1).<sup>21-41</sup> Often called severity measures, their evolution highlights the most important practical concern in widespread risk assessment—the data source.

## Early Development of Severity Measures

Researchers began creating tools for systematic severity measurement in the 1970s, generally to address local concerns.<sup>42</sup> For instance, the measure called Computerized Severity Index (CSI) descends from efforts to help a specific hospital respond to state regulators' questions about its resource use.<sup>43</sup> Another measure, MedisGroups, evolved from an

initiative at Saint Vincent Hospital, Worcester, Mass, to document patient experience at that institution.<sup>44</sup> Jefferson Medical College, Philadelphia, Pa, convened 23 physicians to create a measure, Disease Staging, to aid in evaluating California's Medicaid program.<sup>45</sup>

These early efforts shared one important feature: few data were available to guide their development. Therefore, they were largely normative, built on expert judgment or observed clinical practice at specific institutions. They relied on a "medical model" of acute illness, information typically used by physicians in evaluating specific patients—diagnoses, acute physiological findings, and routine test results. MedisGroups typified this approach. Two Saint Vincent physicians observed morning report of the medical residents, noting clinical parameters that drove residents' severity assessments.<sup>30,44</sup> These observations led to MedisGroups' initial list of key clinical findings (KCFs).

Medicare's adoption of DRG-based prospective payment in 1983 elevated severity concerns to national prominence.<sup>46</sup> While DRG categories initially were de-

rived clinically, their developers made a strategic choice to use data that were routinely available in computerized hospital discharge abstracts.<sup>47</sup> Thus, empirical refinements of the DRGs used more than 1.5 million discharge abstracts and statistical techniques to group case patients with similar lengths of stay.

Discharge abstracts are produced by hospitals on the Medicare claim (UB-92) and for all hospitalizations in states that require them for administrative purposes (eg, rate setting). Discharge abstracts include patient demographic data; payer information; principal and other diagnoses and procedures coded using the *International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)*; admission source; and discharge disposition (eg, death). Diagnosis codes in discharge abstracts represent conditions treated throughout the hospitalization, whether present on admission or arising subsequently. Since diagnosis codes largely determine DRG assignment and thus hospital reimbursement, concerns grew about the possibility of "DRG creep"—the coding of diagnoses not supported by clinical documentation.<sup>48</sup> In addition, studies sug-

gested that important risk factors, especially chronic illnesses, are undercoded.<sup>49-51</sup> Nonetheless, despite limited clinical information of questionable accuracy, discharge abstracts offer the important advantages of uniformity, widespread availability, and computer readability.

Recognizing these benefits, developers of some severity measures (eg, Disease Staging<sup>39,40</sup> and Patient Management Categories [PMCs] Severity Scale<sup>52</sup>) translated their severity logic into ICD-9-CM codes and designed software applicable to computerized discharge abstracts. While much clinical detail was lost, at least these methods competed with the DRGs in being applicable to readily available, large-scale data sets. Furthermore, Disease Staging and PMCs used empirical analyses on large discharge abstract files to refine the specifications of their method.

### Quality Concerns and the Next Generation of Severity Measures

Payment based on the DRG heightened concerns about quality of care, which included fears that hospitals would skimp on inpatient services and discharge patients "quicker and sicker." Using their Medicare claims files, the Health Care Financing Administration (HCFA) had examined hospital mortality rates; these data were released publicly only after a Freedom of Information Act request from the *New York Times*.<sup>7</sup> HCFA's initial publication of hospital-level mortality figures in March 1986 proved problematic.<sup>53</sup> The facility with the most aberrant death rate (87.6% compared with HCFA's prediction of 22.5%) was a hospice caring for terminally ill patients. This early HCFA model had serious methodological flaws, including inadequate severity adjustment.<sup>54</sup>

At the same time, leaders in several states and regions wanted proof about the value of local hospital care, especially given the staggering increases in hospital costs.<sup>7,14</sup> An early effort was the "buy right" experiment in Minneapolis, Minn, led by Walter McClure, who viewed clinically credible data as essential to valid risk adjustment. McClure argued, "There is no other way than to go in and abstract the clinical findings from the chart. Let's stop fooling ourselves that we can compare patient severity by claims."<sup>55</sup> The "buy right" program contracted with MedisGroups, abstracting hundreds of KCFs from medical records to measure severity.

Building off the experience in Minneapolis, the state legislature in Pennsylvania adopted what became known as Act 89 in 1986, largely at the urging of local businesses. Section 3 of Act 89 defined severity as "in any patient, the measurable degree of the potential for failure of one or more vital organs"—MedisGroups' definition. Pennsylvania required all hospi-

tals to report information abstracted from medical records using MedisGroups. Soon afterward, Iowa and Colorado also required most hospitals to gather MedisGroups information. With data now arriving from hundreds of hospitals, the developers of MedisGroups compiled a large, computerized file containing KCF information. They used this "MedisGroups comparative database" to develop and test refinements to their method.

Thus, in the late 1980s, initiatives to risk adjust hospital mortality rates often required clinical information from medical records for severity adjustment. In addition to the MedisGroups mandates in Pennsylvania, Iowa, and Colorado, other initiatives also concentrated on collecting clinical data for risk adjustment, including the Cleveland Health Quality Choice program,<sup>17</sup> a similar effort in St Louis, the Department of Veterans Affairs Surgical Risk Study,<sup>56</sup> and programs to evaluate coronary artery bypass graft (CABG) mortality in New York State<sup>31-34</sup> and northern New England.<sup>55,56</sup> With growing clinical databases, empirical techniques could now be used to refine severity measures. In addition, with relatively little effort, different models could be derived empirically to predict different outcomes (eg, different MedisGroups versions to predict in-hospital death or length of stay). This new generation of severity measures, therefore, appeared both scientifically rigorous and clinically credible.

Nevertheless, the costs of clinical data remained an impediment. One study estimated that Pennsylvania's data collection costs averaged \$17.43 per case patient, with total costs per hospital ranging from \$70 000 at small rural facilities to \$134 000 at large urban hospitals.<sup>8</sup> Pennsylvania businesses, such as Hershey Foods, countered that they saved millions by negotiating with hospitals based on the data. In contrast, Iowa halted collection of MedisGroups data in 1994, arguing that the cost was not worth the information benefit. Iowa chose a discharge abstract-based severity measure called All Patient Refined DRGs (APR-DRGs) for risk adjustment. Colorado's MedisGroups program, dissolved in 1995, was also replaced by APR-DRGs.<sup>57</sup> While electronic clinical data systems may lessen data acquisition costs in the future, widespread availability of such clinical information across hospitals is unlikely any time soon.

California's experience crystallized the debate over data sources. In 1990, after hearing Hershey Foods representatives laud Pennsylvania's MedisGroups initiative, California business leaders proposed legislation to enact a similar program. However, the bill died because of its projected \$61.2 million annual price tag. In 1991, California's legislature mandated production of

"home-grown" risk-adjusted outcome measures, using the state's existing discharge abstract database.<sup>16</sup> A similar experience in Florida ended with use of discharge abstract-based APR-DRGs.

However, concerns about the credibility and accuracy of discharge abstract-based measures persisted. In June 1993, newly appointed HCFA administrator Bruce Vladeck discontinued publication of the Medicare hospital mortality reports, citing the inadequacy of HCFA's discharge abstract-based risk-adjustment method, especially for inner-city public hospitals.<sup>7</sup> Questions about data quality forced California to conduct a special study of data accuracy, which found striking variations across hospitals in the validity and reliability of coding certain risk factors.<sup>13</sup> Variation in coding accuracy explained part of the differences between hospitals that were viewed as high- and low-mortality hospitals; overcoding (coding conditions not supported by medical record documentation) rates ranged from 10% at a putatively high-mortality hospital to 74% at a facility considered low mortality.<sup>13</sup>

Thus, current initiatives to compare risk-adjusted outcomes across hospitals sort into 2 camps: those willing and those unwilling to pay for abstracting clinical data from medical records. The decision about which path to choose is local, with the first camp touting clinical credibility (and thus, presumably, power to sway physicians) and the second arguing resource constraints.

### HOW DO SEVERITY MEASURES COMPARE?

Choosing among severity measures is often daunting (Table 1), but selecting different approaches produces a patchwork quilt of competing methods. In Ohio alone, Cleveland uses their home-grown clinical data-based method,<sup>17</sup> Cincinnati uses Iameter's discharge abstract-based Acuity Index Method (AIM),<sup>9</sup> and the Dayton Employer Coalition recently chose the vendor of the Disease Staging severity measure.<sup>58</sup> Given the differences across severity measures, might different measures give different results?

Despite their widespread use, role in report cards, and potential influence in competitive health care marketplaces, severity measures have received little external scrutiny. Most publications about specific severity methods come from developers, and statistical performance measures vary (Table 2).<sup>59-63</sup> Because patient samples and statistical methods differ, comparing results across studies is problematic. Few comparative studies have been performed,<sup>59-62</sup> and most report on versions of the severity measures that are now out-of-date (Table 2). Conducting comparative studies is hampered by the absence

Table 2.—Examples of Statistical Performance of Severity Measures for Predicting Hospital Deaths\*

Severity Measure	Source, y	Patient Sample	Statistical Measures Reported
<b>Clinical Data—Based Methods</b>			
APACHE III	Knaus et al, <sup>23</sup> 1991 Rosenthal and Harper, <sup>17</sup> 1994	17 440 Intensive care unit (ICU) admissions, 40 hospitals 70 000 ICU admissions, 31 Cleveland-area hospitals	c=0.90† ROC‡ area=0.90
CHOC	Rosenthal and Harper, <sup>17</sup> 1994	31 Cleveland-area hospitals Acute myocardial infarction (AMI) Congestive heart failure Pneumonia or chronic obstructive pulmonary disease Stroke Coronary artery bypass graft (CABG)	ROC area=0.882 ROC area=0.837 ROC area=0.898 ROC area=0.881 ROC area=0.813
CSI	Horn et al, <sup>25</sup> 1991  Alemi et al, <sup>41</sup> 1990§ MacKenzie et al, <sup>42</sup> 1991§	5 Hospitals Pneumonia (n=220) AMI (n=229) 355 Medically treated AMI patients 2955 Medicare patients	R=.469, c=0.861 R=.682, c=0.904 ROC area=0.81 ROC area=0.7275¶
MMPS	Daley et al, <sup>27</sup> 1988	Medicare beneficiaries ≥65 y Stroke (n=1639) Pneumonia (n=1496) AMI (n=1774) Congestive heart failure (n=1650)	Cross-validated R <sup>2</sup> R <sup>2</sup> =0.253 R <sup>2</sup> =0.181 R <sup>2</sup> =0.190 R <sup>2</sup> =0.123
MedisGroups	Steen et al, <sup>28</sup> 1993	MedisGroups Comparative Database Ischemic heart disease (n=47 109) Heart failure (n=22 337) Bacterial lung infection (n=6736) Other lung infection (n=19 449)	Validation sample c=0.877 c=0.788 c=0.837 c=0.874
CSRS	Hannan et al, <sup>22</sup> 1994 Chassin et al, <sup>34</sup> 1996 Green and Winfield, <sup>63</sup> 1995	57 187 New York CABG cases, 1989-1992 New York CABG cases, 1992 New York CABG cases	c=0.787 c=0.826 R <sup>2</sup> =7.3
NNECVDSG	O'Connor et al, <sup>26</sup> 1992	3035 CABG cases, 5 New England centers	ROC area=0.76
<b>Discharge Abstract—Based Methods</b>			
AIM	MacKenzie et al, <sup>42</sup> 1991§	3547 Medicare beneficiaries	ROC area=0.7345¶
California	Wilson et al, <sup>13</sup> 1996	AMI discharges, California hospitals Model A: only diagnoses present on admission No prior admissions within 8 wks (n=62 570) Prior admission within 8 wks (n=5442) Model B: all diagnoses No prior admissions within 8 wks (n=62 220) Prior admission within 8 wks (n=5415)	c=0.774 c=0.759 c=0.860 c=0.830
Disease Staging	MacKenzie et al, <sup>42</sup> 1991§   Alemi et al, <sup>41</sup> 1990§	3560 Medicare patients 355 Medically treated AMI patients	ROC area=0.7374¶ ROC area=0.67
PMCs Severity Scale	Young et al, <sup>41</sup> 1994  Alemi et al, <sup>41</sup> 1990§ MacKenzie et al, <sup>42</sup> 1991§	State discharge abstract databases Maryland 1988 and 1990  California 1990  355 Medically treated AMI patients 3553 Medicare patients	R=0.62 to 0.67 c=0.67 to 0.71 R=0.65 c=0.76 ROC area=0.72 ROC area=0.7508¶

\*All studies examined in-hospital deaths except as noted (by ||). See footnote in Table 1 for expansion of severity measure abbreviations.

†The c statistic measures how well severity measure models discriminate between patients who lived and those who died.

‡ROC area indicates area under a receiver operating characteristic curve; equivalent to c.

§The version of the severity measure used in this study is now largely out-of-date.

||Examined deaths within 30 days of hospital admission.

¶Model also adjusted for adjacent diagnosis related group.

of a single database containing all data elements in the exact form required by different severity measures and the proprietary ownership of many measures.

Our research group has conducted the only extensive, external study of several commercial severity measures sold for comparing hospital mortality rates.<sup>64-73</sup> We studied over a dozen approaches, but for brevity, I summarize findings from 5. Three are discharge abstract-based: Disease Staging,<sup>39,40</sup> PMCs,<sup>41</sup> and APR-DRGs.<sup>37,38</sup> The other 2 use clinical data abstracted from medical records: the admission MedisGroups score<sup>28,29</sup> and physiology scores patterned after the acute physiology component of the third version of the Acute Physiology and Chronic Health Evaluation (APACHE III).<sup>23,24</sup> The physiology scores used 17 physiological values (eg, blood pressure, pulse, se-

rum sodium, hematocrit, arterial pH, etc) drawn from admission MedisGroups KCFs. Physiology scores were studied not to test APACHE per se but to include a method based on a subset of variables that might be more feasible to implement than MedisGroups. APACHE acute physiology scores could not be replicated exactly because specific values for MedisGroups KCFs were not collected in broadly defined normal ranges (eg, MedisGroups only gathered specific systolic blood pressure values ≤90 mm Hg).<sup>74</sup>

The database, described in detail elsewhere,<sup>64-71</sup> was drawn from the 1992 MedisGroups Comparative Database, which contained all standard discharge abstract and KCF data on calendar year 1991 discharges from 108 acute care hospitals around the country. Within each condition, institutions with fewer than 30 pa-

tients were eliminated. Admission MedisGroups scores were contained in the database. Scores for the 3 discharge abstract-based methods were assigned by their vendors, using computer files containing the necessary discharge abstract data elements drawn by us from the MedisGroups database (the database included up to 20 ICD-9-CM diagnosis and 50 procedure codes). Using admission KCF values (eg, systolic blood pressure), physiology scores were created by assigning weights specified by APACHE III and summing them to produce scores.<sup>64-71</sup>

All patients were at least 18 years old. Reported here are 4 conditions with sufficient numbers of in-hospital deaths for meaningful statistical analysis, as follows: medically treated acute myocardial infarction (AMI)—11 880 patients from 100 hospitals with 1574 in-hospital deaths

Table 3.—Statistical Performance of Severity Measures\*

Severity Measures	Acute Myocardial Infarction		Coronary Artery Bypass Graft		Pneumonia		Stroke	
	c (95% CI)	R <sup>2</sup> (95% CI)	c (95% CI)	R <sup>2</sup> (95% CI)	c (95% CI)	R <sup>2</sup> (95% CI)	c (95% CI)	R <sup>2</sup> (95% CI)
MedisGroups	0.83 (.83-.85)	22.7 (21.1-24.7)	0.73 (.70-.76)	3.6 (2.4-5.4)	0.85 (.85-.86)	19.0 (17.9-21.0)	0.87 (.86-.88)	26.5 (24.3-29.1)
Physiology score	0.83 (.82-.84)	22.9 (21.0-24.8)	0.72 (.70-.76)	2.8 (1.7-5.3)	0.81 (.81-.83)	14.9 (14.0-16.7)	0.84 (.83-.85)	24.2 (21.8-26.7)
Disease Staging	0.86 (.85-.87)	27.0 (24.8-28.9)	0.77 (.74-.80)	6.9 (4.3-9.8)	0.80 (.80-.82)	13.2 (12.5-14.8)	0.74 (.73-.76)	11.2 (9.2-13.3)
PMC Severity Scale	0.82 (.81-.83)	17.6 (16.3-19.4)	0.80 (.78-.83)	7.9 (5.9-11.1)	0.79 (.79-.80)	11.5 (11.0-12.8)	0.73 (.72-.75)	10.1 (8.6-11.8)
APR-DRGs	0.84 (.83-.85)	19.8 (18.4-21.4)	0.83 (.81-.86)	6.6 (5.5-8.3)	0.78 (.78-.80)	10.1 (9.9-12.0)	0.77 (.75-.78)	10.5 (9.0-12.6)

\*Conditions and statistical performance measures, c and R<sup>2</sup> × 100. CI indicates confidence interval. See footnote in Table 1 for expansion of severity measure abbreviations.

(13.2%)<sup>64,65</sup>; CABG surgery—7765 patients from 38 hospitals with 252 in-hospital deaths (3.2%)<sup>65,66</sup>; pneumonia—18016 patients from 105 hospitals with 1732 in-hospital deaths (9.6%)<sup>66,70</sup>; and medically treated stroke—9407 patients from 94 hospitals with 916 in-hospital deaths (9.7%)<sup>67,71</sup>.

Logistic regressions were run within each condition to predict death using patient age, sex, and severity scores.<sup>64-71</sup> Separate models were produced for each severity measure, and their statistical performance was judged using the c statistic (which measures how well a severity measure discriminates between patients who lived and those who died, so that c=0.5 indicates no ability to discriminate, while c=1.0, perfect discrimination)<sup>75,76</sup> and R<sup>2</sup>. Cross-validation demonstrated that the models were not overfit. Eighty bootstrap replications<sup>77</sup> of each model were performed to put 95% confidence intervals around the c and R<sup>2</sup> values. Each model produced a predicted probability of in-hospital death for each patient, which was added to determine expected death rates for each hospital. A z score was calculated for each hospital as follows: z = (observed number of deaths - expected number of deaths) / (SD in the number of deaths). Hospitals were ranked from lowest (fewer deaths than expected) to highest (more deaths than expected) based on these z scores.

#### Predicting In-Hospital Death for Individual Patients

**Statistical Performance.**—Potential purchasers of severity measures often look first at the statistical performance of severity measures summarized by c and R<sup>2</sup> statistics. As shown in Table 3, our c and R<sup>2</sup> values were similar to others reported in the literature (Table 2), and no severity measure performed best across all conditions.<sup>64-71</sup> MedisGroups produced the highest c and R<sup>2</sup> for pneumonia and stroke; Disease Staging yielded the highest c and R<sup>2</sup> for AMI; while APR-DRGs had the highest c and PMCs the highest R<sup>2</sup> for CABG. Thus, a clinically based measure performed best for pneumonia and stroke, while discharge abstract-based measures did best for AMI and CABG.

Given the limited clinical information available to discharge abstract-based measures, their superior performance is surprising, until one remembers that they use discharge diagnoses—codes representing all conditions treated during the hospitalization, regardless of when they occurred. In contrast, both clinically based measures used findings from the admission period, which was the first 2 hospital days, up to the time of surgery. We explored whether discharge abstract-based measures rely on diagnosis codes for events occurring late during hospitalization (eg, cardiac arrest).<sup>68-70</sup> Although preliminary, our results are similar to those of others,<sup>13,78</sup> suggesting that the predictive ability of discharge abstract-based measures comes partly from codes for late, "near death" events—post hoc assessments that obviously improve predictive ability but compromise their validity for making inferences about quality.

Another Table 3 result is noteworthy: the similar performance of MedisGroups and physiology scores. Their c statistics were identical for AMI, and the maximum difference between the 2 was 0.04 (for pneumonia). This similarity is striking given their respective origins. MedisGroups scores are based on disease-specific logistic regressions drawing, initially, from up to 250 KCFs; our physiology scores used weights derived for intensive care unit patients, regardless of condition, for only 17 physiological variables. Prior work indicates that physiology scores' predictive performance would improve if they also were derived empirically within specific conditions.<sup>79</sup> Implementing the physiology score would certainly be less expensive than MedisGroups.

**Comparisons of In-Hospital Mortality Predictions.**—The next question was whether different severity measures predict different likelihoods of in-hospital death for the same patients? There are a variety of ways to flag patients viewed as having different predicted probabilities of death<sup>80-71</sup>; however, regardless of the approach, the answer to this question was yes: different severity measures rated many patients very differently. For the analysis presented here, odds ratios (ORs) of death calculated by each severity measure were compared (eg, the odds of dy-

ing predicted by MedisGroups was divided by the odds predicted by Disease Staging).<sup>69,71</sup> When this OR was less than 0.5 or greater than 2.0, a patient was viewed as having a very different probability of death predicted by the 2 measures.

Patients often had dissimilar predicted ORs for dying (Table 4). Again, patterns varied across conditions. For example, agreement between MedisGroups and the physiology scores was high for CABG, but predictions diverged for almost one third of pneumonia patients. Not surprisingly, discharge abstract-based measures disagreed frequently with clinically based measures, but discharge abstract-based measures also differed among themselves. For instance, PMCs and APR-DRGs disagreed on 60.7% of CABG patients.

#### Comparing Hospital Mortality Rates

Finding differences across severity measures at the level of individual patients led to the question: would judgments about whether hospitals look particularly good or bad differ using different severity measures for risk adjustment? The answer to this question is yes—sometimes.<sup>44,47</sup> As before, there are several different approaches to addressing this question. Here, severity-adjusted death rates were used the way report cards or health insurers may use the information. With the z scores, hospitals were identified at 2 extremes:

1. The best 10%: hospitals with the lowest severity-adjusted death rates (the lowest z scores). These hospitals could be designated as exemplar benchmarks for quality improvement initiatives.
2. The worst 10%: hospitals with the highest severity-adjusted death rates (the highest z scores). Insurers may choose not to contract with these institutions.

Table 5 shows how often pairs of severity measures agreed about which hospitals fell into the best and worst 10% and how hospital rankings based on z scores associated with "raw" mortality rates (unadjusted for age, sex, or severity) agree. The immediate impression is that severity measures often flagged different hospitals than unadjusted mortality rates, but different severity measures also frequently flagged different hospitals. No clear pattern emerged, which suggests that discharge abstract-based measures

agreed more with each other than they agreed with the clinically based measures; nor was agreement better for flagging the best 10% than for the worst 10%. Again, results differed by condition (eg, MedisGroups and PMCs agreed relatively well for pneumonia but poorly for stroke).

When disagreements occurred, hospitals ranked in the top or bottom 10% by one severity measure often appeared in the next decile (11% to 20% or 81% to 90%) ranked by the other measure. For example, MedisGroups and the physiology score agreed on 6 of 11 hospitals ranked among the top 10% for pneumonia. The remaining 5, ranked in descending order by MedisGroups (1=best), had the following rankings by the physiology score (MedisGroups rankings are in parentheses): 12 (3), 25 (5), 34 (7), 18 (9), and 14 (10).

Sometimes, however, differences in rankings were large. For instance, MedisGroups and Disease Staging agreed on 5 of 11 hospitals ranked among the top 10% for pneumonia. The remaining 6, ranked in descending order by MedisGroups, had the following rankings by Disease Staging (MedisGroups rankings are in parentheses): 57 (4), 66 (7), 25 (8), 27 (9), 43 (10), and 30 (11). Hospitals with widely discrepant rankings were not low-volume facilities. For instance, the hospital ranked 7th by MedisGroups but 66th by Disease Staging had 266 pneumonia patients with 20 deaths.

Interestingly, agreement in flagging hospitals between severity-adjusted and unadjusted mortality rates was often better than agreement between pairs of severity measures.<sup>64-67</sup> For example, MedisGroups and Disease Staging agreed on only 3 of the 10 worst hospitals for AMI, while MedisGroups and unadjusted rankings agreed on 6 hospitals. Therefore, one set of severity-adjusted findings was not obviously better than another or than unadjusted rankings. For each pair of severity measures,  $\kappa$  statistics were calculated based on whether individual hospitals were flagged as among the best or worst 10% by one, both, or neither measure.<sup>64-67</sup> The  $\kappa$  assesses whether agreement is greater than expected by chance.<sup>60</sup> The  $\kappa$  values showed fair to excellent agreement among severity measures in flagging hospitals (Table 5;  $\kappa$  values are not presented for CABG because the small sample size makes  $\kappa$  less informative).

### Implications

Overall, these analyses suggest that individual hospitals could care greatly about which mortality rates are examined—unadjusted vs severity-adjusted rates—and about which severity measure is used. While severity measures agreed about flagging hospitals more often than chance, rankings for some hospitals differed dramatically across severity measures. A

Table 4.—Percentage of Patients With Different Probabilities of Death Calculated by Pairs of Severity Measures\*

Severity Measures	% of Patients With Different Predicted Odds of Dying			
	Acute Myocardial Infarction	Coronary Artery Bypass Graft	Pneumonia	Stroke
MedisGroups				
Physiology score	19.5	4.1	30.2	17.8
Disease Staging	51.4	32.8	47.6	57.8
PMC Severity Scale	45.6	42.0	46.9	61.6
APR-DRGs	46.1	65.8	47.9	48.4
Physiology score				
Disease Staging	51.6	32.3	38.9	52.0
PMC Severity Scale	44.2	40.9	30.4	44.0
APR-DRGs	44.9	68.5	32.0	31.0
Disease Staging				
PMC Severity Scale	51.8	40.2	31.0	32.9
APR-DRGs	49.5	56.4	30.4	41.3
PMC Severity Scale				
APR-DRGs	27.5	60.7	21.2	20.1

\*Comparisons were based on the equation ratio=(odds of death predicted by first severity measure)/(odds of death predicted by second severity measure). If this ratio was less than 0.5 or greater than 2.0, then the odds predicted by the 2 severity measures were viewed as different. See footnote in Table 1 for expansion of severity measure abbreviations.

Table 5.—Number of Times Pairs of Severity Measures Agreed on the 10% of Hospitals With the Best and Worst Mortality Performance\*

Severity Measures (Including Unadjusted Mortality Rates)	Conditions and No. of Hospitals in Best and Worst 10%							
	Acute Myocardial Infarction† (n=10)		Coronary Artery Bypass Graft‡ (n=4)		Pneumonia§ (n=11)		Stroke   (n=9)	
	Best	Worst	Best	Worst	Best	Worst	Best	Worst
MedisGroups								
Physiology score	9	10	4	4	6	8	7	6
Disease Staging	6	3	2	3	5	5	5	3
PMC Severity Scale	7	5	2	2	6	8	4	3
APR-DRGs	6	4	1	2	6	7	6	4
Unadjusted rates	5	6	3	4	3	6	3	5
Physiology score								
Disease Staging	5	3	2	3	4	7	5	3
PMC Severity Scale	6	5	2	2	7	9	6	3
APR-DRGs	6	4	1	2	5	8	6	3
Unadjusted rates	6	6	3	4	3	9	5	5
Disease Staging								
PMC Severity Scale	7	5	3	2	4	7	6	3
APR-DRGs	7	5	2	2	4	6	7	3
Unadjusted rates	5	4	1	3	1	7	5	3
PMC Severity Scale								
APR-DRGs	8	9	2	2	7	10	7	7
Unadjusted rates	4	6	1	2	5	7	7	6
APR-DRGs								
Unadjusted rates	6	6	1	2	5	7	6	5

\*See footnote in Table 1 for expansion of severity measure abbreviations.

†Number of hospitals on which pairs of severity measures agreed and associated  $\kappa$  value: 3,  $\kappa=0.22$ ; 4,  $\kappa=0.33$ ; 5,  $\kappa=0.44$ ; 6,  $\kappa=0.56$ ; 7,  $\kappa=0.67$ ; 8,  $\kappa=0.78$ ; 9,  $\kappa=0.89$ ; and 10,  $\kappa=1.00$ .

‡Because of small number of hospitals,  $\kappa$  values were unreliable.

§Number of hospitals on which pairs of severity measures agreed and associated  $\kappa$ : 5,  $\kappa=0.39$ ; 6,  $\kappa=0.49$ ; 7,  $\kappa=0.59$ ; 8,  $\kappa=0.70$ ; 9,  $\kappa=0.80$ ; and 10,  $\kappa=0.90$ .

||Number of hospitals on which pairs of severity measures agreed and associated  $\kappa$ : 3,  $\kappa=0.26$ ; 4,  $\kappa=0.39$ ; 5,  $\kappa=0.51$ ; 6,  $\kappa=0.63$ ; and 7,  $\kappa=0.75$ .

logical, albeit worrisome, conclusion is that wise hospitals should "shop around" for the severity measure that shows them to the best advantage, although this measure could vary by condition.

Choosing among different severity measures requires consideration of a variety of factors. Despite their statistical performance, compelling reasons argue against discharge abstract-based mea-

asures for quality measurement initiatives intending to affect physicians' behavior. Convincing clinicians that discharge abstract-based measures (regardless of their statistical performance) are meaningful is hampered by long-standing concerns about ICD-9-CM,<sup>51</sup> coding completeness,<sup>49-51</sup> and financial motivation for diagnosis coding.<sup>48</sup> Additionally, to draw meaningful conclusions about quality

based on severity-adjusted death rates, one must adjust only for preexisting conditions but not those arising late in the hospital stay—possibly because of iatrogenic complications. Our preliminary analyses agree with others<sup>13,35</sup> in suggesting that discharge abstract-based measures rely on late events to boost their predictive ability. This raises the potential for “death code creep”—increased coding of catastrophic events for dying patients—as a hospital response to discharge abstract-based report cards.

#### WHAT DO RISK-ADJUSTED HOSPITAL DEATH RATES MEAN?

Hospitals vary in their unadjusted death rates. According to our work and the findings of others,<sup>35,32,33</sup> severity fails to explain these differences fully. As shown in California, differences across hospitals in accuracy of data can account partially for discrepancies in risk-adjusted death rates, using measures based on discharge abstracts.<sup>13</sup> The central question remains unresolved: does severity adjustment isolate that residual quantity, namely quality-of-care differences, across hospitals?<sup>34</sup> A definitive answer is unlikely any time soon. The research required is expensive, time-consuming, logistically difficult, and methodologically complicated. Only a handful of studies have addressed this question, and most have produced equivocal results.

In an early study, Dubois et al<sup>34</sup> found no systematic differences in quality assessed using process of care measures between high- and low-mortality outlier hospitals; using subjective reviews by expert clinicians, however, more preventable deaths were identified at facilities with higher-than-expected death rates. Using clinically detailed severity<sup>4</sup> and quality measures, Kahn et al<sup>35</sup> found that risk-adjusted mortality rates were significantly related to explicit judgments about process of quality of care for 4 of 5 conditions studied. Using PMCs for risk adjustment, Thomas et al<sup>36</sup> found that hospital mortality performance was significantly related to quality of care for only 3 of 10 conditions evaluated. Another study compared HCFA's hospital mortality ratings with quality problems determined by peer review organizations in 38 states; in 14, confirmed problem rates were statistically significantly correlated with adjusted mortality rates.<sup>37</sup> New York's CABG program found significantly more quality-of-care problems among deaths at high-mortality rather than low-mortality outlier hospitals.<sup>31</sup>

Other studies failed to find any relationship between risk-adjusted mortality rates and hospital quality. Using the same rigorous severity and quality measures as Kahn et al,<sup>35</sup> Park et al<sup>38</sup> found that hospitals flagged as having unexpectedly high

age-, sex-, race-, or disease-specific death rates did not have worse quality than other hospitals. Another study found no association between quality of care and observed-to-expected mortality ratios calculated by the US Department of Veterans Affairs using discharge abstract-type data.<sup>39</sup>

#### WHY RISK ADJUST?

Report cards comparing death rates across hospitals are unlikely to vanish. Nonetheless, uncertainty persists about what risk-adjusted hospital death rates really mean. In addition, research suggests that different risk adjusters can produce different judgments about a hospital's mortality performance.<sup>64-67</sup> No studies are available, or likely, to tell us which risk adjuster is best, especially for isolating quality differences. The initial question, therefore, returns: given these vagaries, why risk adjust? The major reason is that, however imperfect, there is no other way to begin a productive dialogue with physicians and other clinicians about using outcomes information to motivate quality improvement.

Another compelling reason to risk adjust is to avoid penalizing providers who treat high-risk patients. As report cards naming individual hospitals and physicians increasingly make the front pages of newspapers, anecdotal evidence suggests that some physicians and hospitals turn away patients they view as especially high risk.<sup>30,31</sup> Ensuring clinically credible risk adjustment could guard against this. Most current severity measures, however, do not include all patient characteristics that increase risk, such as physical functional status, patients' preferences for care and outcomes, cultural factors, and socioeconomic characteristics.<sup>19</sup> Risk must be assessed within subpopulations of patients (eg, racial and ethnic minorities, medically indigent persons) without bias. Even with clinically derived risk adjustment, some hospitals, like the Cleveland Clinic, will argue that their patients are different.<sup>32</sup>

Designing a clinically reasonable but logistically feasible risk adjustment method is challenging and demands trade-offs. In today's highly charged, competitive environment, criticisms of risk adjustment are inevitable and often appropriate. The accumulated evidence suggests that strong inferences about quality should not be made based on risk-adjusted mortality rates alone. However, the risk of not risk adjusting is that information—albeit imperfect—will be summarily dismissed. Opportunities will be lost for stimulating the introspection required to improve quality of care.

This article was prepared in collaboration with the Center for Studying Health System Change, Washington, DC, with funding from the Robert Wood Johnson Foundation, Princeton, NJ. The

project was supported by the Agency for Health Care Policy and Research, Rockville, Md, under grant No. RO1 HS06742.

The 4-year severity study could not have happened without the creativity, energy, and determination of Michael Schwartz, PhD, Arlene S. Ash, PhD, Yevgenia D. Mackierman, Jennifer Daley, MD, and John S. Hughes, MD. Bruce Landon, MD, also provided valued insight.

#### References

1. Iezzoni LI. 100 Apples divided by 15 red herrings: a cautionary tale from the mid-19th century on comparing hospital mortality rates. *Ann Intern Med.* 1996;124:1079-1085.
2. Response to letter by William Farr. *Med Times Gazette.* February 13, 1864:187. Letter.
3. Epstein A. Performance reports on quality. *N Engl J Med.* 1996;333:57-61.
4. Keeler EB, Kahn KL, Draper D, et al. Changes in sickness at admission following the introduction of the prospective payment system. *JAMA.* 1990;264:1962-1968.
5. Iezzoni LI, Schwartz M, Moskowitz MA, Ash AS, Sawitz E, Burnside S. Illness severity and costs of admissions at teaching and nonteaching hospitals. *JAMA.* 1990;264:1426-1431.
6. 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.
7. US General Accounting Office, Health, Education, and Human Services Division. *Employers and Individual Consumers Want Additional Information on Quality.* Washington, DC: US General Accounting Office; 1995. GAO/HEHS-95-201.
8. US General Accounting Office, Health, Education, and Human Services Division. *Health Care Reform: 'Report Cards' Are Useful but Significant Issues Need to Be Addressed.* Washington, DC: US General Accounting Office; 1994. GAO/HEHS-94-219.
9. US General Accounting Office, Health, Education, and Human Services Division. *Employers Urge Hospitals to Battle Costs Using Performance Data Systems.* Washington, DC: US General Accounting Office; 1994. GAO/HEHS-95-1.
10. Pennsylvania Health Care Cost Containment Council. *A Consumer Guide to Coronary Artery Bypass Graft Surgery, Volume IV: 1993 Data.* Harrisburg: Pennsylvania Health Care Cost Containment Council; 1995.
11. Pennsylvania Health Care Cost Containment Council. *Focus on Heart Attack in Western Pennsylvania: A 1993 Summary Report for Health Benefits Purchasers, Health Care Providers, Policymakers, and Consumers.* Harrisburg: Pennsylvania Health Care Cost Containment Council; 1996.
12. Wilson P, Smoley SR, Werdegar D. *Annual Report of the California Hospital Outcomes Project.* Sacramento, Calif: Office of Statewide Health Planning and Development; 1993.
13. Wilson P, Smoley SR, Werdegar D. *Second Report of the California Hospital Outcomes Project: Acute Myocardial Infarction, Volume One: Study Overview and Results Summary.* Sacramento, Calif: Office of Statewide Health Planning and Development; 1996.
14. Iezzoni LI, Schwartz M, Restuccia J. The role of severity information in health policy debates. *Inquiry.* 1991;28:117-128.
15. Iezzoni LI, Greenberg LG. Widespread assessment of risk-adjusted outcomes: lessons from local initiatives. *Jt Comm J Qual Improv.* 1994;20:305-316.
16. Romano PS, Zach A, Luft HS, Rainwater J, Remy LL, Campa D. The California hospital outcomes project. *Jt Comm J Qual Improv.* 1995;21:668-682.
17. Rosenthal GE, Harper DL. Cleveland health quality choice: a model for collaborative community-based outcomes assessment. *Jt Comm J Qual Improv.* 1994;20:425-442.
18. 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.

19. Iezzoni LI. Dimensions of risk. In: Iezzoni LI, ed. *Risk Adjustment for Measuring Healthcare Outcomes*. 2nd ed. Chicago, Ill: Health Administration Press; 1997.
20. Calkins DR, Rubenstein LV, Cleary PD, et al. Failure of physicians to recognize functional disability in ambulatory patients. *Ann Intern Med*. 1991; 114:451-454.
21. Kahn KL, Pearson ML, Harrison ER, et al. Health care for black and poor hospitalized Medicare patients. *JAMA*. 1994;15:1169-1174.
22. Burstin HR, Lipsitz SR, Brennan TA. Socioeconomic status and risk for substandard medical care. *JAMA*. 1992;268:2383-2387.
23. 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.
24. 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.
25. Horn SD, Sharkey PD, Buckle JM, Backofen JE, Averill RF, Horn RA. The relationship between severity of illness and hospital length of stay and mortality. *Med Care*. 1991;29:305-317.
26. Iezzoni LI, Daley J. A description and clinical assessment of the computerized severity index. *Qual Rev Bull*. 1992;18:44-52.
27. Daley J, Jencks S, Draper D, Lenhart G, Thomas N, Walker J. Predicting hospital-associated mortality for Medicare patients. *JAMA*. 1988;260: 3617-3624.
28. 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.
29. Steen PM. Approaches to predictive modeling. *Ann Thorac Surg*. 1994;58:1836-1840.
30. Iezzoni LI, Moskowitz MA. A clinical assessment of MedisGroups. *JAMA*. 1988;260:3159-3163.
31. Hannan EL, Kilburn H, O'Donnell JF, Lukacik G, Shields EP. Adult open heart surgery in New York State: an analysis of risk factors and hospital mortality rates. *JAMA*. 1990;264:2768-2774.
32. Hannan EL, Kilburn H, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. *JAMA*. 1994;271:761-766.
33. Hannan EL, Siu AL, Kumar D, Kilburn H Jr, Chassin MR. The decline in coronary artery bypass graft surgery mortality in New York State. *JAMA*. 1995;273:209-213.
34. Chassin MR, Hannan EL, DeBuono BA. Benefits and hazards of reporting medical outcomes publicly. *N Engl J Med*. 1996;334:394-398.
35. O'Connor GT, Plume SK, Olmstead EM, et al, for the Northern New England Cardiovascular Disease Study Group. A regional prospective study of in-hospital mortality associated with coronary artery bypass grafting. *JAMA*. 1991;266:803-809.
36. O'Connor GT, Plume SK, Olmstead EM, et al, for the Northern New England Cardiovascular Disease Study Group. Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery from the Northern New England Cardiovascular Disease Study Group. *Circulation*. 1992;35:2110-2118.
37. Edwards N, Honemann D, Burley D, Navarro M. Refinement of the Medicare diagnosis-related groups to incorporate a measure of severity. *Health Care Financing Rev*. 1994;16:45-64.
38. Goldfield N, Boland P, eds. *Physician Profiling and Risk Adjustment*. Gaithersburg, Md: Aspen Publishers Inc; 1996.
39. Gonnella JS, Hornbrook MC, Louis DZ. Staging of disease: a case-mix measurement. *JAMA*. 1984; 251:637-644.
40. Markson LE, Nash DB, Louis DZ, Gonnella JS. Clinical outcomes management and Disease Staging. *Eval Health Prof*. 1991;14:201-227.
41. Young WW, Kohler S, Kowalski J. PMC patient severity scale: derivation and validation. *Health Serv Res*. 1994;29:367-390.
42. Iezzoni LI. Severity of illness measures and assessing the quality of hospital care. In: Goldfield N, Nash DB, eds. *Providing Quality Care: Future Challenges*. 2nd ed. Ann Arbor, Mich: Health Administration Press; 1995:59-82.
43. Horn SD. Validity, reliability and implications of an index of inpatient severity of illness. *Med Care*. 1981;19:354-362.
44. Brewster AC, Karlin BG, Hyde LA, et al. MEDISGRPS: a clinically based approach to classifying hospital patients at admission. *Inquiry*. 1985;12: 377-387.
45. Gonnella JS, Louis DZ, McCord JJ. The staging concept—approach to the assessment of outcome of ambulatory care. *Med Care*. 1976;14:13-21.
46. Vladeck BC. Medicare hospital payment by diagnosis-related groups. *Ann Intern Med*. 1984;100: 576-591.
47. Fetter RB, Shin Y, Freeman JH, Averill R, Thompson J. Case mix definition by diagnosis related groups. *Med Care*. 1980;18(suppl):1-53.
48. Simborg DW. DRG creep: a new hospital-acquired disease. *N Engl J Med*. 1981;304:1602-1604.
49. Jencks SF, Williams DK, Kay TL. Assessing hospital-associated deaths from discharge data. *JAMA*. 1988;260:2240-2246.
50. 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.
51. Green J, Wintfeld N. How accurate are hospital discharge data for evaluating effectiveness of care? *Med Care*. 1993;31:719-731.
52. Young WW, Swinkola RB, Zorn DM. The measurement of hospital case mix. *Med Care*. 1982;20: 501-512.
53. Brinkley J. US releasing lists of hospitals with abnormal mortality. *New York Times*. March 12, 1986:1.
54. Dubois RW. Hospital mortality as an indicator of quality. In: Goldfield N, Nash DB, eds. *Providing Quality Care: Future Challenges*. 2nd ed. Ann Arbor, Mich: Health Administration Press; 1995.
55. Iglehart JK. Competition and the pursuit of quality: a conversation with Walter McClure. *Health Aff (Millwood)*. 1988;7:79-90.
56. Khuri SF, Daley J, Henderson WG, et al. The National Veterans Administration Surgical Risk Study: risk adjustment for the comparative assessment of the quality of surgical care. *J Am Coll Surg*. 1995;180:519-531.
57. Demise of two state-run commissions signals shift to voluntary initiatives in data collection. *State Health Watch*. 1995;2:4, 10.
58. Verna G. Dayton hospitals link to perform cost study. *Cincinnati Business Courier*. 1996;13:8C.
59. Thomas JW, Ashcraft MLF. Measuring severity of illness: a comparison of interrater reliability among severity methodologies. *Inquiry*. 1989;26:483-492.
60. Thomas JW, Ashcraft MLF. Measuring severity of illness: six severity systems and their ability to explain cost variations. *Inquiry*. 1991;28:39-55.
61. Alemi F, Rice J, Hanks R. Predicting in-hospital survival of myocardial infarction. *Med Care*. 1990;28:762-775.
62. MacKenzie TA, Willan AR, Lichter J, et al. *Patient Classification Systems: An Evaluation of the State of the Art*. Kingston, Ontario: Case Mix Research, Queen's College; 1991:1.
63. Green J, Wintfeld N. Report cards on cardiac surgeons: assessing New York State's approach. *N Engl J Med*. 1995;332:1229-1232.
64. Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Judging hospitals by severity-adjusted mortality rates. *Am J Public Health*. 1996; 86:1379-1387.
65. Landon B, Iezzoni LI, Ash AS, et al. Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. *Inquiry*. 1996;33:155-166.
66. Iezzoni LI, Shwartz M, Ash AS, Hughes JS, Daley J, Mackiernan YD. Severity measurement methods and judging hospital death rates for pneumonia. *Med Care*. 1996;34:11-28.
67. 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.
68. Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Predicting who dies depends on how severity is measured. *Ann Intern Med*. 1995;123: 763-770.
69. Iezzoni LI, Ash AS, Shwartz M, Landon B, Mackiernan YD. Predicting in-hospital deaths from CABG surgery. *Med Care*. In press.
70. Iezzoni LI, Shwartz M, Ash AS, Mackiernan YD. Using severity measures to predict the likelihood of death for pneumonia patients. *J Gen Intern Med*. 1996;11:23-31.
71. Iezzoni LI, Shwartz M, Ash AS, Mackiernan YD. Predicting in-hospital mortality for stroke patients: results differ across severity measurement systems. *Med Decis Making*. 1996;16:348-356.
72. Hughes JS, Iezzoni LI, Daley J, Greenberg L. How severity measures rate hospitalized patients. *J Gen Intern Med*. 1996;11:303-311.
73. Iezzoni LI, Ash AS, Shwartz M, Mackiernan YD. Differences in procedure use, outcomes, and illness severity by gender for acute myocardial infarction patients. *Med Care*. 1997;35:153-171.
74. 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.
75. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med*. 1984;3:143-152.
76. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29-36.
77. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7:1-26.
78. Hannan EL, Racz MJ, Jollis JG, Peterson ED. Using Medicare claims data to assess provider quality for CABG surgery: does it work well enough? *Health Serv Res*. 1997;31:659-678.
79. Iezzoni LI, Ash AS, Coffman GA, Moskowitz MA. Predicting in-hospital mortality: a comparison of severity measurement approaches. *Med Care*. 1992;30:347-359.
80. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
81. McMahon LF, Smits HL. Can Medicare prospective payment survive the ICD-9-CM disease classification system? *Ann Intern Med*. 1986;104:562-566.
82. Dubois RW, Brook RH, Rogers WH. Adjusted hospital death rates. *Am J Pub Health*. 1987;77: 1162-1167.
83. Green J, Passman LJ, Wintfeld N. Analyzing hospital mortality. *JAMA*. 1991;265:1849-1853.
84. Dubois RW, Rogers WH, Moxley JH, Draper D, Brook RH. Hospital inpatient mortality: is it a predictor of quality? *N Engl J Med*. 1987;317:1674-1680.
85. Kahn KL, Rogers WH, Rubenstein LV, et al. Measuring quality of care with explicit process criteria before and after implementation of the DRG-based prospective payment system. *JAMA*. 1990; 264:1969-1973.
86. Thomas JW, Holloway JJ, Guire KE. Validating risk-adjusted mortality as an indicator for quality of care. *Inquiry*. 1993;30:6-22.
87. 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.
88. Park RE, Brook RH, Koseoff J, et al. Explaining variations in hospital death rates, randomness, severity of illness, quality of care. *JAMA*. 1990;264:484-490.
89. Best WR, Cowper DC. The ratio of observed-to-expected mortality as a quality of care indicator in non-surgical VA patients. *Med Care*. 1994;32:390-400.
90. Schneider EC, Epstein AM. Influence of cardiac surgery performance reports on referral practices and access to care. *N Engl J Med*. 1996;335:251-256.
91. Omoigui NA, Miller DP, Brown KJ, et al. Out-migration for coronary bypass surgery in an era of public dissemination of clinical outcomes. *Circulation*. 1996;93:27-33.
92. Vogel RA, Topol EJ. Practice guidelines and physician scorecards: grading the graders. *Cleve Clin J Med*. 1996;63:124-128.



# Hospital Council of Western Pennsylvania

500 Commonwealth Drive • Warrendale, PA 15086-7513 • (724) 776-6400 • (800) 704-8434 • FAX (724) 776-6969  
www.hcwp.org

February 15, 1999

Marc P. Volavka  
Executive Director  
Health Care Cost Containment Council  
Suite 400  
225 Market Street  
Harrisburg, PA 17101

**SENT VIA: Fax and FedEx - 2/15/99**

ORIGINAL: 1995  
MIZNER  
COPIES: de Bien  
Harris  
Sandusky  
Legal

RECEIVED  
99 FEB 19 AM 9:05  
HOSPITAL COUNCIL OF WESTERN PENNSYLVANIA

Dear Marc:

The Hospital Council of Western Pennsylvania (HCWP) is an association representing 38 acute care hospitals in western Pennsylvania. Hospital Council is a key data, strategic planning and advocacy resource for western Pennsylvania's acute care and specialty hospitals, long-term care facilities and rehabilitation centers. Throughout its 60 years of service, this non-profit organization has earned a national reputation for its leadership and excellence in managing health care issues.

We appreciate the opportunity to comment on the proposed amendments to 28 PA. CODE CHS. 911 and 912, which remove specific reference to a particular methodology currently used by the Pennsylvania Health Care Cost Containment Council (the Council). These amendments will afford the Council flexibility in selecting an alternative methodology for measuring provider quality and provider service effectiveness.

In general, we are quite supportive of these regulatory changes and we appreciate the efforts of the Council to address a very significant and costly issue affecting all Pennsylvania hospitals. These changes provide will enable the Council to take actions that can result in an immediate savings to Pennsylvania hospitals well in excess of \$40 million per year. Pennsylvania will also be in a position to adopt a severity system comparable to those used by most other states.

In this letter, we are providing our recommendations on the following sections of the proposed rule making: the statement of purpose, the fiscal impact statement, and the definition of patient severity.

Since 1988, the mandate to use the MediQual system has directly cost Pennsylvania hospitals well in excess of \$400 million in data collection costs alone. For this amount of money, one digit from 0 to 4, indicating patient severity, has been added to each medical record in the PHC4 database.

As you know, the two states which had previously mandated the MediQual system have rescinded that mandate, due to the excess costs related to reporting requirements of the MediQual system and to the availability of high quality alternative systems.

Independent published research has shown that less costly alternative severity systems exist that provide severity measures of a comparable quality, but with a greater breadth than is provided by MediQual. Currently, severity adjusted data analysis of hospital care provided in Pennsylvania can not be legitimately compared to similar studies in other states, due to differences in the way severity scores are determined. The only comparative studies that can be made are among Pennsylvania hospitals, or to the relatively few individual hospitals outside Pennsylvania that have voluntarily elected to participate in the MediQual system.

**AmeriQual**

*Affiliated Entities:*

Hospital Shared Services  
Administrative Resources, Inc.  
Healthcare Information Corp.  
Health Knowledge Systems, Inc.

*Enhancing the capability of the continuum of care to improve the health of communities.*

Marc P. Volavka, Executive Director  
Health Care Cost Containment Council  
February 15, 1999  
Page 2

In order to remove any constraining ambiguity from the proposed regulations, we recommend the following:

1. Under the statement of *Purpose*, our recommendation is that the second sentence be modified as follows:

**"In addition, the proposed amendments will enable the Council to change its vendor if the vendor fails to meet its contractual requirements or if the Council, in consideration of other factors, determines that a change of vendor is appropriate."**

Our recommendation makes it clear that the Council is free to change vendors based on consideration of a variety of significant factors and irrespective of a given vendor's performance.

2. Under the *Fiscal Impact* statement, we take issue with the statement that these changes will have no fiscal impact on the regulated community, which includes every hospital in the Commonwealth. Although there will be no direct fiscal impact due to the adoption of these regulations, the potential indirect fiscal impact on the regulated community will be substantial. If the Council were to select the alternative vendor recommended by its Technical Advisory Committee, the savings to Pennsylvania hospitals would exceed \$40 million per year.

3. Under Annex A, Title 28. HEALTH AND SAFETY, PART VI. HEALTH CARE COST CONTAINMENT COUNCIL, CHAPTER 911. DATA SUBMISSION AND COLLECTION, Subchapter A. STATEMENT OF POLICY, section 911.1 Definitions, the following changes are recommended:

*Patient severity*

In order to provide the Council with the ability to give fair consideration to discharge abstract-based severity systems and not be limited by regulation to only consideration of clinical data-based severity systems, the following definition of patient severity is recommended:

***Patient severity* - A measure of severity of illness as defined by the Council [using appropriate] and determined through the application of either: 1) a reputable discharge abstract-based severity system using appropriate diagnosis, treatment and demographic indicators from the current standard discharge abstract form, or 2) a reputable clinical data-based severity system using appropriate clinical findings such as physician examinations, radiology findings, laboratory findings and pathology findings or any other relevant clinical factors.**

We feel these changes will provide the Pennsylvania Health Care Cost Containment Council with the latitude they need to give fair consideration to alternative severity systems.

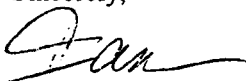
The interpretation of the definition as it is currently presented in the proposed rulemaking will limit consideration of the Council to only those severity systems that require collection of specific clinical findings and indicators. Since this has tremendous cost implications for the regulated community, this is unacceptable and truly undermines the stated purpose of the amendments. To understand why, a background discussion is provided in the attached appendix.

Marc P. Volavka, Executive Director  
Health Care Cost Containment Council  
February 15, 1999  
Page 3

These recommended changes provide the Council with the opportunity to improve the data and analyses it provides. At the same time, these changes demonstrate that the Council is truly committed to health care cost containment. There is no decision or action that the Council could take that would help contain Pennsylvania hospitals' costs now and in the future more than the elimination of the current mandate and the adoption of a discharge abstract-based severity system.

Thank you for your consideration of these important issues. If I or my staff can provide any additional information related to these comments, please call me.

Sincerely,



Ian G. Rawson, Ph. D.  
President

Attachment

c: Carolyn Scanlan, The Hospital & Healthsystem Association of Pennsylvania  
Andrew Wigglesworth, Delaware Valley Hospital Council of HAP  
The Honorable Dennis O'Brien, Representative, State of Pennsylvania  
The Honorable Frank Oliver, Representative, State of Pennsylvania  
The Honorable Harold Mowery, Jr., Senator, State Of Pennsylvania  
The Honorable Vincent J. Hughes, Senator, State of Pennsylvania  
John R. McGinley, Jr., Independent Regulatory Review Commission

## Appendix Background on Severity Systems

Severity systems are typically divided into two categories based upon the origin of the data they require. The first category is termed "discharge abstract-based systems". The second category is termed "clinical data-based systems."

An extensive body of independent published research, as well as some non-published research conducted by specific states, has identified differences in the severity findings and has focused on strengths and weaknesses of both categories of severity systems. The conclusion of this extensive body of research is that, for types of patients addressed by both systems, one category of severity system can not be considered superior or more accurate than the other. For adult medical and surgical patients, the results obtained from the application of a clinical data-based severity system are neither better nor worse than the results obtained from the application of a discharge abstract data-based severity system. After studying this issue, these conclusions were reflected in the recommendations of the Council's Technical Advisory Committee.

However, as is also reflected in independent research literature, discharge abstract-based severity systems provide much wider applicability to other types of hospital patients. Specifically, the discharge abstract-based severity system recommended by the Council's Technical Advisory Committee, is significantly better than the MediQual system in differentiating the severity among pediatric, rehabilitation and psychiatric patients.

Given the above conclusions, other factors must then be considered by the Council in selecting or in mandating a severity system for use. Personal preference or familiarity by the Council and its staff is certainly one consideration. Another consideration should be the comparable cost differences between clinical data-based severity systems and discharge abstract-based severity systems, not only to the Council but also to the regulated community. A third consideration could be the trend observed across the country for states to move away from clinical data-based severity systems and gravitate toward discharge abstract-based systems.

As previously mentioned, the second consideration, specifically that of costs to the regulated hospital community, is a very important issue to the hospital community. The reason for this concern is rooted in the inherent cost of providing the required data for a clinical data-based severity system versus the cost of doing so for a discharge abstract-based severity system.

A clinical severity system, like the MediQual system, requires extensive data collection and abstraction of clinical indicators for each patient covered by the system. A discharge abstract-based severity system requires the submission of data that is currently being collected for billing purposes on every patient. This data is already being provided directly to the Council. The separate collection of clinical data required by the MediQual system costs each hospital between \$20.00 and \$40.00 per patient. This is in addition to the annual software costs of the MediQual system, which are significant.

99 FEB 23 AM 11:55

INDEXED  
MEDICAL  
LITERATURE

NOTICE TO READER: This material may be protected by copyright law.

(Title 17, U. S. Code)

ORIGINAL: 1995

MIZNER

COPIES: de Bien  
Harris  
Sandusky

## Predicting Who Dies Depends on How Severity Is Measured: Implications for Evaluating Patient Outcomes

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

■ **Objective:** To determine whether assessments of illness severity, defined as risk for in-hospital death, varied across four severity measures.

■ **Design:** Retrospective cohort study.

■ **Setting:** 100 hospitals using the MedisGroups severity measure.

■ **Patients:** 11 880 adults managed medically for acute myocardial infarction; 1574 in-hospital deaths (13.2%).

■ **Measurements:** For each patient, probability of death was predicted four times, each time by using patient age and sex and one of four common severity measures: 1) admission MedisGroups scores for probability of death scores; 2) scores based on values for 17 physiologic variables at time of admission; 3) Disease Staging's probability-of-mortality model; and 4) All Patient Refined Diagnosis Related Groups (APR-DRGs). Patients were ranked according to probability of death as predicted by each severity measure, and rankings were compared across measures. The presence or absence of each of six clinical findings considered to indicate poor prognosis in patients with myocardial infarction (congestive heart failure, pulmonary edema, coma, low systolic blood pressure, low left ventricular ejection fraction, and high blood urea nitrogen level) was determined for patients ranked differently by different severity measures.

■ **Results:** MedisGroups and the physiology score gave 94.7% of patients similar rankings. Disease Staging, MedisGroups, and the physiology score gave only 78% of patients similar rankings. MedisGroups and APR-DRGs gave 80% of patients similar rankings. Patients whose illnesses were more severe according to MedisGroups and the physiology score were more likely to have the six clinical findings than were patients whose illnesses were more severe according to Disease Staging and APR-DRGs.

■ **Conclusions:** Some pairs of severity measures assigned very different severity levels to more than 20% of patients. Evaluations of patient outcomes need to be sensitive to the severity measures used for risk adjustment.

*Ann Intern Med.* 1995;123:763-770.

From Harvard Medical School, Beth Israel Hospital, Boston University Medical Center, and Boston University, Boston, Massachusetts; the Brockton/West Roxbury Veterans Affairs Medical Center, West Roxbury, Massachusetts; and the West Haven Veterans Affairs Medical Center, West Haven, Connecticut. For current author addresses, see end of text.

Hospital and physician performance is increasingly scrutinized by organizations ranging from state governments to managed care payers to local business coalitions (1-7). Hospitals and medical practices also monitor their own results to identify areas in which they can produce improvement and savings. Performance profiles of health care providers often compare patient outcomes, such as death rates; comparing such outcomes across hospitals or physicians generally requires adjustment for patient risk. Risk adjustment recognizes that the underlying nature of some patients' diseases makes those patients more likely than others to have poor outcomes (8, 9).

More than a dozen risk-adjustment tools, often called severity measures, have been created specifically to address health care administration and policy concerns (1-7, 10-12). Unlike clinical measures of risk, which can incorporate such factors as disease-specific clinical findings, complexity of comorbid illness, and functional status (13), severity measures rate patients on the basis of limited data—either computerized hospital discharge abstracts (14, 15) or information gathered from medical records by using abstraction protocols independent of specific diseases (16-18). These methods generally focus on predicting hospital resource consumption or in-patient death. They are frequently proprietary, and their complete logic is often unavailable for scrutiny.

Severity measures are now marketed widely to hospitals, payers, business leaders, and governments. Some states (Pennsylvania, Iowa, Colorado, and Florida, for example), regions (such as Cleveland and Orlando), and payers produce comparative performance reports of health care providers by using particular severity measures (1-5). Important decisions are increasingly made on the basis of severity-adjusted patient outcomes. For example, since 1986, Pennsylvania has required hospitals to produce severity information using MedisGroups. Payers have used MedisGroups-based reports to select health care providers for managed care networks (5). Pennsylvania's "consumer guide" (19), which compares hospital death rates and average charges for coronary artery bypass graft surgery, was quoted by President Clinton in his 22 September 1993 health care reform address to the United States Congress (20):

We have evidence that more efficient delivery of health care doesn't decrease quality.... Pennsylvania discovered that patients who were charged \$21,000 for [coronary bypass] surgery received as good or better care [based on MedisGroups severity-adjusted death rates] as patients who were charged \$84,000 for the same procedure in the same state. High prices simply don't always equal good quality.

Despite the potential effects of severity measures, relatively little independent information is available about

Severity Measure	Source	Data Used and Definition of Severity†	Classification Approach	Derivation‡
MedisGroups (18)	MediQual Systems, Inc. (Westborough, Massachusetts)	Clinical data; in-hospital death; score calculated within 64 disease groups	Probability ranging from 0 to 1	Empirical modeling
Physiology score	Patterned after acute physiology score of APACHE III (22, 23)	Clinical data; in-hospital death for patients in intensive care unit	Integer score starting with 0; APACHE III's Acute Physiology Score ranges from 0 to 25‡	Empirical modeling with clinical guidance
Disease Staging (24-26)	Systemetrics/MEDSTAT Group (Santa Barbara, California)	Discharge abstract; probability of in-hospital death	Probability ranging from 0 to 1	Empirical modeling
All Patient Refined Diagnosis Related Groups (27)	3M Health Information Systems (Wallington, Connecticut)	Discharge abstract; total hospital charges	Four severity classes (A, B, C, and D) within adjacent diagnosis-related groups‡	Empirical modeling with clinical guidance

\* APACHE III = Acute Physiology and Chronic Health Evaluation III.  
 † Discharge abstract = standard hospital-discharge data elements, such as basic demographics and diagnosis and procedure codes. Clinical data = clinical information, such as vital signs and test results, abstracted from the medical record.

‡ "Derivation" indicates the principal method used to create the severity scoring method. "Clinical guidance" reflects primarily the use of expert physician guidance; "empirical modeling" indicates primarily the use of statistical techniques.

‡ Adjacent diagnosis-related groups were formed by grouping individual diagnosis-related groups previously split by complications and comorbidities.

them (21). Because they are used to evaluate hospitals and physicians, physicians must assess them, especially with respect to their clinical credibility. In this article, we focus on predicting in-hospital death using four severity measures, and we ask three major questions: 1) How well do severity measures predict in-hospital death? 2) Do different severity measures predict different likelihoods of death for the same patients? and 3) If so, what are the clinical characteristics of patients for whom very different likelihoods of death are predicted by different severity measures?

## Methods

### Severity Measures

We considered four severity measures (Table 1): the admission MedisGroups score (18); a physiology score patterned after the acute physiology score of the Acute Physiology and Chronic Health Evaluation, third version (APACHE III) (22, 23); Disease Staging's scale predicting probability of in-hospital death (24-26); and All Patient Refined Diagnosis Related Groups (APR-DRGs) (27). These systems are among the most prominent approaches used to adjust outcomes data for severity so that they can be used for state or regional comparisons across hospitals (1-5) and for hospital activities such as internal monitoring, negotiation of managed care contracts, and physician profiling.

Each measure defines severity in ways that reflect that measure's goals, assigning either numerical severity scores or values on a continuous scale (Table 1). Disease Staging and APR-DRGs use data from standard hospital-discharge abstracts (14, 15), including patient age, patient sex, and diagnoses and procedures coded using the *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM). A discharge abstract contains codes for all diagnoses treated during a particular hospitalization, regardless of when the diagnoses were made. MedisGroups and the physiology score use clinical data abstracted from medical records for only the first 2 days of a hospitalization.

Although the APR-DRGs measure was not initially developed to predict mortality, it is used for such analyses. For example, Iowa once required larger hospitals to produce MedisGroups data for severity-adjusted performance reports, but it switched in 1984 to using APR-DRGs—a less expensive, discharge-abstract-based measure. This change was partially motivated by the perceived high cost of MedisGroups medical record reviews. Other

states, such as Florida, also use APR-DRGs to evaluate health care provider performance.

### Database

To assign severity scores to patients, computerized algorithms were applied to data extracted from the 1992 MedisGroups Comparative Database. Briefly, this database contains the clinical information collected using the MedisGroups data-gathering protocol and submitted to MedisGroups' vendor, MediQual Systems, Inc. The 1992 MedisGroups Comparative Database contains information on all discharges made in 1991 from 108 acute care hospitals, which were chosen by MediQual Systems because of good data quality and in order to encompass a range of hospital characteristics.

To ensure adequate sample sizes for hospital-level analyses in another study (28), we eliminated eight low-volume institutions (83 patients total). The American Hospital Association annual survey provided information on hospital characteristics.

Admission MedisGroups scores were provided by MediQual Systems; scores for other measures had to be assigned. The MedisGroups database contains standard discharge-abstract information, including ICD-9-CM codes for as many as 20 diagnoses and 50 procedures, listed by hospital. It also includes values for key clinical findings from the admission period (generally the first 2 days of hospitalization), abstracted from medical records during MedisGroups reviews (16-18). We used these clinical findings to create physiology scores patterned after the APACHE III acute physiology score, summing weights specified by APACHE III for each finding (for example, a pulse of 145 beats/min had a weight of 13 points) (22). We could not replicate exact APACHE III acute physiology scores because complete values for the required 17 physiologic variables were unavailable: MedisGroups truncates data collection in broadly defined normal ranges (29). Previous research (29) showed that a similarly derived physiology score did well compared with the exact acute physiology scores of the second APACHE version.

Vendors scored the data for the two discharge-abstract-based severity measures (Table 1). On the basis of their specifications, vendors were sent computer files containing the required discharge-abstract data extracted from the MedisGroups database. We merged the scored data into a single analytic file with 100% success.

### Study Sample and Outcome Measure

Many internal hospital monitoring programs and external evaluations, such as Pennsylvania's MedisGroups initiative (19), sample patients by diagnosis-related groups. To parallel this ap-

proach, we selected all patients in the database who had been hospitalized for medical treatment of a new acute myocardial infarction defined by diagnosis-related groups. We chose acute myocardial infarction because it is a common condition, is treated at most hospitals, and has a relatively high mortality rate. We included patients in diagnosis-related groups 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). Patients had either a principal or secondary 3-digit ICD-9-CM discharge diagnosis code beginning with "410" and ending with "1" (initial treatment).

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

### Analysis

Each severity measure was used to calculate a predicted probability of death for each patient from a multivariable logistic regression model that included 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  $\geq 85$  years of age). Severity scores were entered as either continuous or categorical variables (Table 1). For Disease Staging and MedisGroups, we used the logit of the probability as the independent variable in the logistic regression. All analyses were done using the Statistical Analysis System, release 6.08 (SAS Institute, Cary, North Carolina).

### Severity Measure Performance

We used  $c$  and  $R^2$  statistics as overall assessments of each severity measure's ability to predict individual patient death. The  $c$  statistic assesses this ability as follows: When a person who has died and a person who has lived are each chosen at random,  $c$  equals the probability that the severity measure predicts a higher likelihood of death for the one who has died (30). Higher  $c$  values indicate better specificity and sensitivity (31, 32). A  $c$  value of 0.5 indicates that the model does no better than random chance; a  $c$  value of 1.0 shows perfect performance. The  $R^2$  statistic is commonly interpreted as the percentage of variation in outcomes explained by the model. It is typically lower for models of dichotomous outcomes (such as death) than for models with continuous outcomes (such as length of stay). The  $R^2$  statistic adds independent information to that contributed by the  $c$  statistic for assessing how well predictions match actual outcomes (33).

Sometimes assessments of model performance are overly optimistic when the same data are used to both develop and evaluate models. To guard against this, we calculated cross-validated performance measures ( $c$  and  $R^2$ ) as follows (34): 1) we randomly split the data in half; 2) we estimated coefficients for each model on the first half of the data and calculated "validated" performance measures by applying these coefficients to the second half; and 3) we repeated this process, developing the model on the second half of the data and validating it on the first half. Cross-validated  $c$  and  $R^2$  statistics represent the average of the two validated statistics calculated on the two halves of the data.

For each severity measure, we ranked patients according to their predicted probability of death on the basis of the multivariable model. We then divided patients into 10 groups of equal size (deciles 1 to 10) according to predicted likelihood of dying, and we compared actual and predicted death rates within each decile. These figures suggest how well models separated patients with very high and very low risks for death (model calibration). We also computed a Hosmer-Lemeshow chi-square statistic (35), which measures differences between actual and predicted numbers of deaths within the 10 deciles; goodness of fit is tested by comparing this statistic to a chi-square distribution with 8 degrees of freedom. Given our large sample size, even small differences between observed and expected numbers of deaths were statistically significant.

### Ranking Patients by Predicted Probability of Death

We created  $10 \times 10$  tables, arraying patients within deciles computed by one severity measure against patients within deciles

computed by a second severity measure. The four severity measures thus yielded six  $10 \times 10$  tables, one for each of the six pairwise comparisons. The average probabilities of death for patients within each decile indicated that a difference of three or more deciles constituted an important difference in the predicted likelihood of dying. For each  $10 \times 10$  table, we counted the fractions of patients who had 1) "similar" predicted likelihoods of dying (probabilities of death calculated by severity measures A and B were within two deciles of each other); and 2) "different" predicted likelihoods of dying (probabilities of death calculated by severity measures A and B were three or more deciles apart).

We separated patients with "different" predicted likelihoods of dying into two groups: those for whom the probabilities calculated by severity measure A were much higher than the probabilities calculated by severity measure B; and those for whom the probabilities calculated by severity measure A were much lower than the probabilities calculated by severity measure B. Conceptually, the former group represents patients viewed as "more sick" by measure A than by measure B, and the latter group represents patients viewed as "less sick" by measure A than by measure B.

### Testing Clinical Validity

After finding that different severity measures resulted in very different rankings of the predicted likelihood of dying for the same patients, our next question was, Which severity measure correlated better with clinical findings thought to represent severe illness in patients with acute myocardial infarction? As a preliminary examination of this question, we reviewed the literature on predicting imminent death from myocardial infarction (36-43). We selected six important clinical findings identified in the first 2 days of hospitalization: congestive heart failure, pulmonary edema, coma, low systolic blood pressure ( $\leq 60$  mm Hg), low left ventricular ejection fraction ( $\leq 0.35$ ), and elevated blood urea nitrogen level ( $\geq 31$  mg/dL).

We examined each clinical finding individually for its relation to in-hospital death by creating  $2 \times 2$  tables (finding present/absent by dead/alive) and calculating chi-square statistics. We also computed two logistic regression models: Both had dummy variables for each clinical finding, and one also included age and sex categories. We report odds ratios with 95% CIs for death for each clinical finding. Results are given  $\pm$  SD.

We counted the percentage of patients with each clinical finding among persons with different predicted likelihoods of dying for each pair of severity measures.

### Analyses of ICD-9-CM Complication Codes

The predictive ability of discharge-abstract-based measures could relate to the fact that these measures consider all discharge diagnosis codes, regardless of whether the conditions coded for were present at admission or developed subsequently. To examine whether discharge-abstract-based measures relied heavily on conditions developing after admission, we used ICD-9-CM codes to define serious conditions, including cardiac arrest, respiratory arrest, respiratory failure, and coma. We did analyses identical to those done with the six clinical findings (see above) using these conditions defined by ICD-9-CM codes.

### Results

The final data set contained 11 880 patients and 1574 in-hospital deaths (13.2%). Patients ranged from 19 to 103 years of age (mean,  $68.3 \pm 13.3$  years); 58.1% of patients were men. For Disease Staging and APR-DRGs, ample numbers of diagnosis codes were usually present for rating severity (mean,  $5.6 \pm 3.0$  diagnosis codes per patient). Only 4.2% of patients had 1 discharge diagnosis code; 43.4% had more than 5 diagnoses listed; and 10.2% had 10 or more diagnoses listed.

Fifty-five of the 100 hospitals were in Pennsylvania, and 16 were from the southern United States. The 100 hospitals were generally larger, more likely to offer cardiac intensive care, more likely to be urban, less likely to be

**Table 2. Severity Measure Performance for Predicting Death: Statistical Fit on All Cases and Cross-Validated Statistics\***

Statistical Performance Measure	Severity Measure			
	MedisGroups	Physiology Score	Disease Staging DRGs	APR-DRGs
c statistic	0.834	0.832	0.862	0.842
Cross-validated c statistic	0.833	0.831	0.861	0.841
R <sup>2</sup> statistic	0.228	0.230	0.270	0.200
Cross-validated R <sup>2</sup> statistic	0.226	0.228	0.268	0.196

\* APR-DRGs = All Patient Refined Diagnostic Related Groups.

public, and more involved in teaching than other general acute care institutions nationwide (28).

#### Statistical Performance

The four severity measures varied in their statistical performance as measured by their *c* and *R*<sup>2</sup> values (Table 2). Cross-validated performance was identical to or only 0.01 points lower than the performance of models developed using the entire data set. The two clinical data-based measures (MedisGroups and the physiology score) had almost identical *c* and *R*<sup>2</sup> values. The two discharge-abstract-based measures (Disease Staging and APR-DRGs) had similar *c* statistics, which were slightly higher than those of the clinical data-based measures.

Table 3 shows the actual and predicted death rates for patients within each of the 10 deciles. All severity measures arrayed patients along wide ranges of predicted probabilities of death. Disease Staging had the broadest range: Patients in the lowest decile had a predicted death rate of 0.4% (actual death rate, 0.3%), and those in the highest decile had a predicted death rate of 59.7% (actual death rate, 58.4%). As measured by the Hosmer-Lemeshow chi-square statistic, MedisGroups had the best calibration: The actual and predicted death rates within each of the 10 deciles were not significantly different (*P* = 0.317). However, given the large sample size, it was not surprising that the Hosmer-Lemeshow values for the other three measures were statistically significant, even when differences between the actual and predicted death rates within deciles appeared small.

**Table 3. Actual and Predicted Death Rates within 10 Deciles\***

Decile	MedisGroups		Physiology Score		Disease Staging		APR-DRGs	
	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
	←----- % -----→							
1	0.5	0.6	0.3	0.7	0.3	0.4	0.0	0.7
2	1.3	1.8	1.3	2.1	0.4	1.2	0.9	2.0
3	2.4	3.1	2.4	3.3	1.5	2.3	1.6	2.8
4	3.8	4.5	4.5	5.1	2.9	3.6	3.0	3.7
5	5.7	6.3	5.4	6.5	4.0	5.3	4.6	4.6
6	9.5	8.5	10.6	8.6	8.4	7.6	8.7	6.2
7	12.1	11.5	12.9	11.2	12.1	10.7	11.7	9.9
8	17.3	16.2	16.5	15.5	17.8	15.9	20.5	20.4
9	25.9	25.4	23.7	23.9	26.6	25.7	36.1	32.6
10	53.9	54.6	54.9	55.7	58.4	59.7	45.4	49.6
<i>P</i> value	0.317		0.004		0.002		0.0001	

\* Deciles were created by ranking patients according to increasing predicted likelihood of dying on the basis of severity score, age, and sex and by dividing them into 10 equal groups. APR-DRG = All Patient Refined Diagnosis Related Groups.

#### Comparison of Patients with Different Predicted Likelihoods of Death

Table 4 shows the percentage of patients with similar and different predicted probabilities of death for pairs of severity measures and the percentage of patients who died within each group. The rankings of MedisGroups and the physiology score were generally similar; these measures assigned similar likelihoods of dying to 94.7% of patients. In contrast, the two discharge-abstract-based measures frequently disagreed with the clinical data-based measures about patient severity, ranking many patients differently according to likelihood of dying. For example, MedisGroups and Disease Staging ranked 22.3% of patients very differently.

Patients viewed as sicker by the discharge-abstract-based measures than by the clinical data-based measures had a higher death rate than did those viewed as less sick by the former measures than by the latter. For example, of the 11.0% of patients viewed as sicker by Disease Staging than by MedisGroups, 16.5% died; in contrast, of the 11.3% of patients viewed as sicker by MedisGroups than by Disease Staging, only 10.8% died.

#### Clinical Validity Analysis

Each of the six clinical findings had a strong individual relation (*P* < 0.0001) with in-hospital death (Table 5), which supports the validity of these findings as indicators of risk for death from acute myocardial infarction. The logistic regression that included age and sex plus the six clinical findings yielded a *c* statistic of 0.81 and an *R*<sup>2</sup> value of 0.21. When controlling for other clinical findings (but not age and sex), the odds ratios for predicting in-hospital death were as follows: congestive heart failure, 1.66 (95% CI, 1.43 to 1.93); pulmonary edema, 1.22 (CI, 1.00 to 1.49); coma, 11.75 (CI, 9.52 to 14.51); low systolic blood pressure, 6.38 (CI, 5.16 to 7.91); low ejection fraction, 1.29 (CI, 1.04 to 1.60); and elevated blood urea nitrogen level, 3.58 (CI, 3.13 to 4.07).

Table 6 shows the percentage of patients with each clinical finding among patients for whom pairs of severity measures produced different likelihoods of dying. For example, 11.2% of patients viewed as sicker by MedisGroups than by Disease Staging had low systolic blood



Table 4. Percentage of Patients with Very Different Relative Predicted Probabilities of Death Calculated by Pairs of Severity Measures and Percentage of These Patients Who Died\*

Severity Measure		Comparison of Predicted Likelihood of Dying by Severity Measure		
A	B	A's Prediction Similar to B's Prediction	A's Prediction Much Higher than B's Prediction	A's Prediction Much Lower than B's Prediction
		← Patients, % (Patients Who Died, %) →		
MedisGroups	Physiology Score	94.7 (13.2)	2.7 (11.7)	2.6 (14.5)
MedisGroups	Disease Staging	77.7 (13.1)	11.5 (10.8)	11.0 (16.5)
MedisGroups	APR-DRGs	79.4 (12.7)	10.2 (13.2)	10.4 (17.2)
Physiology Score	Disease Staging	77.8 (13.2)	11.5 (10.3)	10.7 (16.8)
Physiology Score	APR-DRGs	80.1 (12.6)	9.8 (13.3)	10.1 (18.7)
Disease Staging	APR-DRGs	84.0 (13.8)	6.9 (11.2)	9.1 (9.3)

\* Total patients = 11 890. APR-DRG = All Patient Refined Diagnosis Related Groups.

pressure at admission; in contrast, only 0.8% of patients seen as sicker by Disease Staging than by MedisGroups had low systolic blood pressure. Patients viewed as sicker by MedisGroups were more likely to have each of the six findings than were patients seen as sicker by Disease Staging.

In general, patients viewed as sicker by a clinical data-based measure than by a discharge-abstract-based measure were more likely to have each clinical finding than were patients for whom the opposite was true.

#### Analyses of ICD-9-CM Complication Codes

Most complications defined by ICD-9-CM codes occurred too rarely to allow rigorous statistical analysis. In one exception, 6.0% of patients had cardiac arrest codes; 60.4% of these patients died, compared with 10.2% of patients who did not have cardiac arrest codes ( $P < 0.001$ ). Among patients viewed as sicker by Disease Staging than by MedisGroups, 16.2% had cardiac arrest codes; only 0.4% of patients seen as sicker by MedisGroups than by Disease Staging had cardiac arrest codes. The comparison of Disease Staging and the physiology score produced similar findings. These results, albeit preliminary, support the idea that codes such as that for cardiac arrest play an important role in discharge-abstract-based ratings of patient severity.

#### Discussion

Detailed evaluation of severity measures appears to be a narrow methodologic pursuit, far removed from daily medical practice. Nevertheless, severity-adjusted death rates are widely used as putative quality indicators in health care provider "report cards" (1-7). Because severity measures could significantly affect their practices, physicians should assist—and possibly take the lead—in evaluating the validity of these measures. Examining the clinical credibility of severity measures demands extensive physician input. Physicians should ensure that the methods used to evaluate clinical performance are open to external scrutiny.

Our findings suggest, however, that interpreting such evaluations is complicated. The "take-home" messages of these evaluations may not be a definitive "This measure is good and that is bad." Discharge-abstract-based severity measures (Disease Staging and APR-DRGs) were slightly

better able to predict death (measured by  $c$  and  $R^2$  values) than the clinical data-based measures (MedisGroups and the physiology score). In contrast, MedisGroups and the physiology score had better clinical credibility (relations with six clinical indicators of risk for death from acute myocardial infarction) than Disease Staging and APR-DRGs. Thus, the discharge-abstract-based measures had better predictive validity, and the clinical data-based measures had better clinical validity.

Differences between severity measures that occur on the basis of their data sources have important health policy implications, suggesting a trade-off between data costs and clinical credibility. Because discharge-abstract data are routinely produced by hospitals, they are generally available, computer accessible, and inexpensive. These advantages have led some provider evaluation initiatives around the country (for example, in California, Connecticut, Florida, New Jersey, and Ohio) to use discharge abstracts. As stated earlier, Iowa switched from MedisGroups to APR-DRGs largely because of data costs. Because discharge abstracts often result from billing, however, some investigators have questioned whether financial

Table 5. Frequency of Clinical Findings and Associated Death Rates\*

Clinical Finding at Admission	Patients	In-Hospital Death†	
		Patients with Clinical Finding	Patients without Clinical Finding
← n(%) →			
Congestive heart failure	1804 (15.2)	412 (22.8)	1162 (11.5)
Pulmonary edema	854 (7.2)	216 (25.3)	1358 (12.3)
Coma	503 (6.2)	333 (66.2)	1241 (10.9)
Low systolic blood pressure ( $\leq 60$ mm Hg)	487 (4.1)	259 (53.2)	1315 (11.5)
Low left ventricular ejection fraction ( $\leq 0.35$ )	734 (6.2)	139 (18.9)	1435 (12.9)
Elevated blood urea nitrogen level ( $\geq 31$ mg/dL)	2119 (17.8)	657 (31.0)	917 (9.4)
Any clinical finding	4308 (36.3)	1129 (26.2)	445 (5.9)

\* Total patients = 11 890.

†  $P = 0.0001$  for comparison of death rates of patients with and without each specific clinical finding.

Table 6. Percentage of Patients with Very Different Probabilities of Death Predicted by Different Severity Methods Who Had Specific Clinical Findings\*

Relative Predicted Probability of Death for Pairs of Severity Measures	Clinical Finding at Admission						
	CHF	Pulmonary Edema	Coma	Low SBP	Low LVEF	High BUN	Any
	← Patients, % →						
MedisGroups > Physiology score (n = 324)	15.7	6.8	0.0	12.0	14.5	6.5	42.6
MedisGroups < Physiology score (n = 310)	12.6	8.1	22.3	1.6	6.1	24.2	55.2
MedisGroups > Disease Staging (n = 1337)	21.5	11.2	3.4	11.2	10.1	27.5	55.9
MedisGroups < Disease Staging (n = 1312)	10.2	4.5	2.8	0.8	4.8	8.5	24.1
MedisGroups > APR-DRGs (n = 1209)	16.0	8.9	2.8	10.7	7.8	21.8	46.6
MedisGroups < APR-DRGs (n = 1235)	15.1	7.1	3.0	0.4	6.6	12.1	32.8
Physiology score > Disease Staging (n = 1368)	21.3	11.6	6.1	7.1	7.1	31.3	55.3
Physiology score < Disease Staging (n = 1271)	11.3	4.0	0.0	1.5	7.1	5.0	22.9
Physiology score > APR-DRGs (n = 1159)	16.4	9.3	5.9	8.4	4.9	24.4	47.0
Physiology score < APR-DRGs (n = 1201)	15.7	6.4	0.0	1.8	7.8	8.5	30.3
Disease Staging > APR-DRGs (n = 814)	10.6	4.7	3.7	3.3	5.7	11.9	29.2
Disease Staging < APR-DRGs (n = 1062)	24.2	11.6	3.0	3.3	8.6	23.8	51.5

\*APR-DRG = All Patient Refined Diagnosis-Related Group; CHF = congestive heart failure; Low SBP = systolic blood pressure  $\leq 60$  mm Hg; Low LVEF = left ventricular ejection fraction  $\leq 0.35$ ; High BUN = blood urea nitrogen level  $\geq 31$  mg/dL; Any = any of these six findings.

motivations compromise data accuracy (44-46). In addition, the clinical information contained in discharge abstracts is limited (47). Nonetheless, our finding that the discharge-abstract-based measures were somewhat better able to predict death supports the choice of these measures.

However, the slightly better predictive ability of discharge-abstract-based measures may result from their inclusion of all diagnoses. These measures review all discharge diagnosis codes, including codes for cardiac arrest, respiratory arrest, ventricular fibrillation, and cardiogenic shock, regardless of when these events occurred. Consideration of the codes may also explain why patients viewed as sicker by the discharge-abstract-based measures than by the clinical data-based measures had higher death rates (Table 4).

Groups around the United States are drawing inferences about health care provider quality on the basis of death rates adjusted for severity using discharge-abstract data (1-4). However, this raises obvious concerns: If quality is to be judged by using severity-adjusted death rates, adjustment should consider only preexisting conditions, not those that develop after hospitalization (48, 49). Otherwise, events occurring late in the hospital stay (possibly as a result of poor care) may mask the detection of deaths due to poor quality. Thus, we focused our clinical analysis on findings from the first 2 days of hospitalization. If severity measures are used to judge quality, it may be reasonable to trade some predictive ability for greater clinical credibility.

Nevertheless, another trade-off remains: Abstracting clinical information from medical records is costlier than relying on existing discharge-abstract data. Given cost concerns, one notable finding is the similarity of the MedisGroups and physiology score results. Our physiology score was patterned after the acute physiology score of APACHE III, using only information from the clinical literature (22). We included physiology scores, not to examine APACHE specifically, but because of growing interest in creating "minimum clinical data sets" containing small numbers of clinical variables. Although APACHE weights are one way to use such variables, other ways

exist. The physiology score requires 17 clinical variables, whereas MedisGroups' data abstraction protocol examines more than 200 potential findings, regardless of patients' diagnoses.

Our study has important limitations. We looked at just one condition. The database contained information only from hospitals using MedisGroups; independent information about data reliability was unavailable. The clinical findings were specifically gathered for MedisGroups scoring, possibly giving MedisGroups an advantage in statistical performance and the clinical validity analysis. The MedisGroups algorithm for rating the severity of ischemic heart disease explicitly considers congestive heart failure, coma, low ejection fraction, low systolic blood pressure, and high blood urea nitrogen levels, among many other variables (13). All measures are periodically revised; newer versions may provide different results.

The MedisGroups database contains information on only in-hospital deaths. Knowing mortality rates after discharge permits holding the "window of observation" constant (for example, at 30 days after admission). This is important when comparing patient mortality across health care providers with different discharge practices (50). However, comparing death rates across hospitals was not our goal. We have no reason to expect that our overall finding—that different severity measures ranked many patients differently according to probability of death—would be different if we had looked at 30-day mortality.

Finally, our work is not a comprehensive comparative evaluation of severity systems; a complete study would require attention to additional issues. Commenting on the evaluation of quality measurement methods, Donabedian (51) suggested that "the concept of validity is itself made up of many parts [and] covers two large domains. The first has to do with the accuracy of the data and the precision of the measures that are constructed with these data. The second has to do with the justifiability of the inferences that are drawn from the data and the measurements." Using this conceptual framework, a major remaining challenge is to examine whether judgments made on the basis of severity-adjusted death rates are justified.

Does this information really offer insight into health care provider quality?

Our results suggest that mortality analyses require sensitivity to the severity adjustment measure used. Because different measures often rank the same patients at different severity levels, different hospitals or physicians may be viewed as having particularly good or bad risk-adjusted patient death rates, depending on the severity adjustment measure used. Our findings also raise concern about the use of severity scores (or predictions of imminent death) in making decisions about care for individual patients, because perceptions of the illness severity of individual patients may depend on the specific severity measure used.

Given the potential effects of severity measures on patients and health care providers, a formal process to evaluate them seems justified. Reason suggests that before a method is used to judge health care provider performance, it should be proven to measure quality. In the current health policy environment, however, rules of evidence and proof appear to be reversed. Because they are often the only measures available, severity-adjusted mortality rates will be used as indicators of health care provider quality until someone proves that they are not appropriate for this purpose. A definitive study is unlikely to be done anytime soon. Such research is expensive, and it poses the daunting challenge of defining "gold standard" quality measures. Nevertheless, both the public and health care providers need assurance that the information generated by using severity measures is valid.

**Grant Support:** This research was supported by grant RO1 HS06742-05 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.

**Requests for Reprints:** Lisa I. Iezzoni, MD, Division of General Medicine and Primary Care, Department of Medicine, Beth Israel Hospital, Room LY-326, 330 Brookline Avenue, Boston, MA 02215.

**Current Author Address:** Dr. Iezzoni and Ms. Macklerman: Division of General Medicine and Primary Care, Department of Medicine, Beth Israel Hospital, Room LY-326, 330 Brookline Avenue, Boston, MA 02215. Dr. Ash: Health Care Research Unit, Section of General Internal Medicine, Boston University Medical Center, 720 Harrison Avenue, Room 1108, Boston, MA 02118. Dr. Schwartz: Health Care Management Program and Operations, Management Department, School of Management, Boston University, 621 Commonwealth Avenue, Boston, MA 02215. Dr. Daley: Health Services Research and Development, Department of Medicine, Brockton/West Roxbury Veterans Affairs Medical Center, 1400 VFW Parkway, West Roxbury, MA 02132. Dr. Hughes: Department of Medicine, Department of Veterans Affairs Medical Center, 950 West Campbell Avenue, West Haven, CT 06516.

#### References

1. United States General Accounting Office. Health Care Reform: Report Cards Are Useful but Significant Issues Need to Be Addressed. Report to the Chairman, Committee on Labor and Human Resources, U.S. Senate, Washington, DC: U.S. General Accounting Office; 1994 [GAO/HEHS-94-219].
2. United States General Accounting Office. Health Care: Employers Urge Hospitals to Battle Costs Using Performance Data Systems. Report to Congressional Requesters, Washington, DC: U.S. General Accounting Office; 1994 [GAO/HEHS-95-1].
3. Iezzoni LI, Schwartz M, Resaccio J. The role of severity information in health policy debates: a survey of state and regional concerns. *Inquiry*. 1991;28:117-28.
4. Iezzoni LI, Greenberg LG. Widespread assessment of risk-adjusted outcomes: lessons from local initiatives. *Jt Comm J Qual Improv*. 1994;20:305-16.
5. Localio AR, Hamory BF, 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;122:123-32.
6. Green J, Winfield N. Report card on cardiac surgeons. Assessing New York State's approach. *N Engl J Med*. 1995;332:1229-32.
7. Epstein A. Performance reports on quality—prototypes, problems, and prospects. *N Engl J Med*. 1995;333:37-61.
8. Selker HP. Systems for comparing actual and predicted mortality rates: characteristics to promote cooperation in improving hospital care [Editorial]. *Ann Intern Med*. 1993;118:820-2.
9. Kassirer JP. The use and abuse of practice profiles [Editorial]. *N Engl J Med*. 1994;330:634-6.
10. McMahon LF Jr, Billi JE. Measurement of severity of illness and the Medicare prospective payment system: state of the art and future directions. *J Gen Int Med*. 1988;3:482-90.
11. The Quality Measurement and Management Project. The Hospital Administrator's Guide to Severity Measurement Systems. Chicago: The Hospital Research and Educational Trust of the American Hospital Association; 1989.
12. Iezzoni LI, ed. Risk Adjustment for Measuring Health Care Outcomes. Ann Arbor: Health Administration Pr; 1994.
13. Iezzoni LI. Dimensions of risk. In Iezzoni LI, ed. Risk Adjustment for Measuring Health Care Outcomes. Ann Arbor: Health Administration Pr; 1994:29-118.
14. United States Congress, Office of Technology Assessment. Identifying Health Technologies That Warrant Searching for Evidence. Washington, DC: Office of Technology Assessment, Congress of the United States; 1994 [OTA H 608].
15. United States National Committee on Vital and Health Statistics. Uniform Hospital Discharge Data: Minimum Data Set. Report of the National Committee on Vital and Health Statistics. Hyattsville, MD: U.S. Department of Health, Education, and Welfare, Public Health Service, Office of Health Research, Statistics, and Technology, National Centre for Health Statistics; 1980 [DHWQ Pub. No. (PHS) 80-1157].
16. Brewster AC, Karlin BG, Hyde LA, Jacobs CM, Bradbury EC, Chase YM. MEDISORPS: a clinically based approach to classifying hospital patients at admission. *Inquiry*. 1985;12:77-87.
17. Iezzoni LI, Macklerman MA. A clinical assessment of MedisGroups. *JAMA*. 260:3159-63.
18. Sivas PM, Brewster AC, Bradbury EC, Estabrook R, Young JA. Predicted probabilities of hospital death as a measure of admission severity of illness. *Inquiry*. 1993;20:128-41.
19. Pennsylvania Health Care Cost Containment Council. Coronary Artery Bypass Graft Surgery. Technical Report, Volume II 1991. Harrisburg, PA: February 1994.
20. Clinton B. Health Security: The President's Report to the American People. Washington, DC: The White House Domestic Policy Council; 1993:100.
21. Iezzoni LI. "Black box" medical information systems. A technology needs assessment [Editorial]. *JAMA*. 1991;265:3006-7.
22. Kassirer WA, Wagner DP, Draper EA, Zimmerman JE, Burgeon M, Bartos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*. 1991; 100:1619-36.
23. Kassirer WA, Wagner DP, Zimmerman JE, Draper EA. Variations in mortality and length of stay in intensive care units. *Ann Intern Med*. 1993;118:753-61.
24. Coombs JS, Herbrook MC, Leake DZ. Staging of disease: A case-mix measurement. *JAMA*. 1984;251:637-44.
25. Markson LE, Nash DB, Lewis DZ, Coombs JS. Clinical outcomes management and disease staging. Evaluation and the Health Professions. 1991;14:201-27.
26. Nussbaum JM, LeDonne CL, Krishna 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: 1140-9.
27. All Patient Refined Diagnosis Related Groups. Definition Manual. Wallingford, CT: 3M Health Information Systems; 1993.
28. Iezzoni LI, Schwartz M, Ash AS, Hughes JB, Daley J, Macklerman YD, et al. Evaluating severity adjusters for patient outcome studies. Final report. Boston: Beth Israel Hospital; 1995. Prepared for the Agency for Health Care Policy and Research under grant no. RO1-HS06742.
29. Iezzoni LI, Hotchkiss EK, Ash AS, Schwartz M, Macklerman Y. MedisGroups data bases. The impact of data collection guidelines on predicting in-hospital mortality. *Med Care*. 1993;31:277-83.
30. Farrell FE Jr, Lee KL, Caffee RM, Fryer DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med*. 1984;3:143-52.
31. Hasley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29-36.
32. Hasley JA, McNeil BJ. A method of comparing the area under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148:859-63.
33. Ash AS, Schwartz M. Evaluating the performance of risk-adjustment methods: dichotomous measures. In Iezzoni LI, ed. Risk Adjustment

- for Measuring Health Care Outcomes. Ann Arbor: Health Administration Pr; 1994:313-46.
34. Daley J. Validity of risk-adjustment methods. In: Iezzoni LI, ed. Risk Adjustment for Measuring Health Care Outcomes. Ann Arbor: Health Administration Pr; 1994:254.
  35. Lemenow S, Haasler DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*. 1982;115:92-106.
  36. Daley J, Jencks S, Draper D, Leinhardt G, Thomas N, Walker J. Predicting hospital-associated mortality for Medicare patients. A method for patients with stroke, pneumonia, acute myocardial infarction, and congestive heart failure. *JAMA*. 1988;260:3622-4.
  37. Keeler EB, Kabis KL, Draper D, Sherwood MJ, Rubenstein LV, Rotlach EJ, et al. Changes in sickness at admission following the introduction of the prospective payment system. *JAMA*. 1990;264:1962-8.
  38. Morawa R, Sompson T, Yanquella P, Derrida S, Beaumont H, Skot C. Comparison of two simplified severity scores (SAPS and APACHE II) for patients with acute myocardial infarction. *Crit Care Med*. 1989;17:409-13.
  39. Fomen MW, D'Agostino RB, Mitchell JB, Rosenfeld DM, Coughlin JJ, Schwartz ML, et al. The usefulness of a predictive instrument to reduce inappropriate admissions to the coronary care unit. *Ann Intern Med*. 1980;92(2 Pt 1):238-42.
  40. Selker HP, Griffith JL, D'Agostino RB. A time-insensitive predictive instrument for acute myocardial infarction mortality: a multicenter study. *Med Care*. 1991;29:1196-211.
  41. Teskey RJ, Cahria JE, McPhail I. Disease severity in the coronary care unit. *Chest*. 1991;100:1637-42.
  42. Wagner DP, Knans WA, Draper EA. Physiologic abnormalities and outcome from acute disease. Evidence for a predictable relationship. *Arch Intern Med*. 1986;146:1389-96.
  43. Iezzoni LI, Ash AS, Coffman GA, Moskowitz MA. Predicting in-hospital mortality. A comparison of severity measurement approaches. *Med Care*. 1992;30:347-59.
  44. Vladeck BC. Medicare hospital payment by diagnosis-related groups. *Ann Intern Med*. 1984;100:776-91.
  45. Simborg DW. DRG creep: a new hospital-acquired disease. *N Engl J Med*. 1981;304:1602-4.
  46. Hris DC, Kroschek WM, Fagan AB, Tebbett JA, Kussnerow KP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *N Engl J Med*. 1988;318:352-5.
  47. McMahon LF Jr, Smith BL. Can Medicare prospective payment survive the ICD-9-CM disease classification system? *Ann Intern Med*. 1986;104:562-6.
  48. Iezzoni LI, Foley SM, Horva T, Daley J, Duncan CC, Fisher ES, et al. A method for screening the quality of hospital care using administrative data: preliminary validation results. *QRB Qual Rev Bull*. 1992;18:361-71.
  49. 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;29:93-112.
  50. Jencks SF, Williams DK, Kay TL. Assessing hospital-associated deaths from discharge data. The role of length of stay and comorbidities. *JAMA*. 1988;260:2240-6.
  51. Donabedian A. Explorations in Quality Assessment and Monitoring. 1. The Definition of Quality and Approaches to Its Assessment. Ann Arbor: Health Administration Pr; 1980:101.



THE HOSPITAL & HEALTHSYSTEM ASSOCIATION OF PENNSYLVANIA

RECEIVED

Carolyn F. Scanlan  
President and Chief Executive Officer

99 FEB 19 PM 3: 55

INDEPENDENT REGULATORY  
REVIEW COMMISSION

February 16, 1999

Mr. Marc P. Volavka  
Executive Director  
Pennsylvania Health Care Cost  
Containment Council  
225 Market Street, Suite 400  
Harrisburg, Pennsylvania 17101

ORIGINAL: 1995  
MIZNER  
COPIES: Nyce  
de Bien  
Harris  
Sandusky  
Legal

Re: 28 PA. Code Chapters 911 and 912

Dear Marc:

The Hospital & Healthsystem Association of Pennsylvania (HAP), on behalf of its members (more than 225 acute and specialty hospitals and health systems in the commonwealth), appreciates the opportunity to comment on the Pennsylvania Health Care Cost Containment Council's proposed rulemaking (published in the *Pennsylvania Bulletin* on January 16, 1999) amending the council regulations.

The current regulation specifies a particular methodology to evaluate the effectiveness of patient care. That methodology was selected based on available systems in 1987. By specifying a particular methodology, the council is precluded from selecting a different vendor and/or methodology that may be more effective and economical. As HAP understands the proposed amendments, their purpose is to give the council the flexibility to utilize a different vendor if it appears that a more effective and economical system is available. It also gives the council the opportunity to rapidly seek another vendor and/or methodology if the current vendor (MediQual) fails to perform. Based upon this understanding of the intent of the proposed amendments, HAP supports the proposed rulemaking.

Two relevant issues regarding the proposed rulemaking require specific attention by the council, the Independent Regulatory Review Commission, and the legislature. The council needs flexibility to select a patient severity methodology which allows the council to measure the effectiveness of health care providers. Additionally, the potential impact of the proposed rulemaking needs to be better understood. The following observations/ recommendations address these two issues.

4750 Lindle Road  
P.O. Box 8600  
Harrisburg, PA 17105-8600  
717.564.9200 Phone  
717.561.5334 Fax  
cscanlan@hap2000.org



Mr. Marc P. Volavka  
February 16, 1999  
Page 2

### **Patient Severity**

To ensure that the council has the needed flexibility to select a more effective and economical severity methodology, it is important that the amendments permit the council to consider the full array of severity adjustment systems currently available and which may be developed in the future. The proposed change in the definition of "patient severity" could prove to limit the council's flexibility in selecting an alternative methodology, even if the alternative methodology provides better and more complete information on the effectiveness of health care providers.

**HAP recommends that the council modify the definition of "patient severity" to afford greater flexibility in selecting severity methodologies. HAP recommends the following definition of "patient severity" be used by the council in final rulemaking:**

*Patient severity*—A measure of severity of illness as defined by the council through the application of either: 1) a reputable discharge abstract-based severity system using appropriate indicators (i.e., diagnosis, treatments, demographics, and resource utilization) from the standard patient discharge abstract; or 2) a reputable severity system using data (i.e., diagnosis, treatments, demographics, and other relevant factors) abstracted from individual patient records.

### **Potential Economic or Fiscal Impact**

Adoption of the proposed rulemaking, in and of itself, will have no fiscal impact. The proposed amendments will not, per se, impose additional paperwork requirements. However, when the council chooses to exercise the flexibility afforded it through the proposed rulemaking there is the potential for significant impact (positive and/or negative).

Currently, hospitals incur significant costs and paperwork requirements associated with collecting data using the mandated MediQual system (estimated at \$40 million to \$50 million annual cost for all Pennsylvania hospitals). These costs include the fees paid by hospitals to MediQual to license the mandated severity adjustment system as well as the cost for personnel to manually complete MediQual patient abstract forms for approximately 1.3 million inpatient discharges per year, enter the abstracted data into the proprietary MediQual software, transmit the abstracted data to MediQual, and validate

Mr. Marc P. Volavka  
February 16, 1999  
Page 3

the abstracted data. Additionally, the difficulties encountered in implementing MediQual's Atlas 2.0 software illustrate the potential for unnecessary burdens on hospitals. The costs of the MediQual mandate are in addition to the costs hospitals incur in creating standard patient discharge abstract data sets that are submitted directly to the council for all inpatient discharges and all ambulatory surgery cases.

Two of the council's stated ongoing objectives are "to make data collection more effective and to potentially reduce costs incurred by reporting providers." Consequently, if the council elects to adopt a different methodology and/or vendor, the cost of compliance should be a principal consideration. The council should also consider the benefits to the commonwealth of adopting a different methodology. If an alternative methodology provides better information on the effectiveness of health care providers, all Pennsylvania residents, their insurance companies and/or their employers could make better choices on selecting providers. Improved information on the health care market and potentially lower costs of compliance could spawn a more competitive market which could improve the quality of care at a lower cost.

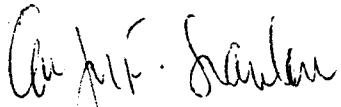
**HAP recommends that following adoption of the final-form publication of revised amendments (see patient severity above) the council exercise due diligence in exploring alternative severity adjustment systems for possible adoption. Due to the potential direct impact on the regulated community, the council should seek involvement of the regulated community in the selection of a severity adjustment methodology.**

In summary, HAP supports the intent of the proposed regulations and believes that they can be improved by enhancing the definition of "patient severity" to allow the council greater flexibility in evaluating severity adjustment methodologies. Additionally, HAP recommends that, upon adoption of the proposed rulemaking, the council exercise due diligence in evaluating severity adjustment systems with the active participation of the regulated community.

Mr. Marc P. Volavka  
February 16, 1999  
Page 4

HAP is committed to improving the timeliness, quality, and effectiveness of data reported to and by the council. We believe that the proposed rulemaking is an important step in reaching this objective. We offer our cooperation and assistance in whatever capacity is needed. If you have any questions, or if we can be of further assistance, please feel free to call me at (717) 561-5314 or Martin Ciccocioppo at (717) 561-5363.

Sincerely,

A handwritten signature in black ink, appearing to read 'Carolyn F. Scanlan', written in a cursive style.

CAROLYN F. SCANLAN  
President and Chief Executive Officer

CFS/mjc

- c: Honorable Dennis O'Brien, Chair, House Health and Human Services Committee
- Honorable Frank Oliver, Minority Chair, House Health and Human Services Committee
- Honorable Harold Mowery, Jr., Chair, Senate Public Health and Welfare Committee
- Honorable Vincent Hughes, Minority Chair, Senate Public Health and Welfare Committee



RECEIVED  
99 FEB 23 AM 11:55

ORIGINAL: 1995  
MIZNER  
COPIES: de Bien  
Harris  
Sandusky

SYNOPSIS OF RELEVANT LITERATURE

---

Predicting In-Hospital Mortality: A comparison of severity measurement approaches.

Medical Care, April 1992, Vol. 30, No. 4; 347 - 359.

Iezzoni, Ash, Coffman, Moskowitz

(Pg. 357, paragraph 4)

...First, clinical data appeared to be substantially more predictive of in-hospital death than administrative information. Second, a potentially economical and relatively powerful predictor of in-hospital mortality for the five conditions under study (stroke, lung cancer, pneumonia, acute myocardial infarction, congestive heart failure) could be achieved using a smaller number of clinical variables (e.g., those used in the APS or out 10 KCF's model). Third, the ability of admission severity findings to predict inpatient survival varied by condition, and although general, physiologic parameters formed the core subset of predictors of inpatient death, disease-specific parameters were also important. As a result, a generic severity model may not be the most powerful approach for predicting death in specific conditions. A productive approach for risk-adjusting, in-hospital mortality figures may involve adding a small subset of variables to a core generic, physiologic subset. Finally, the most efficient use of the data requires condition-specific weighting of even the generic clinical findings. Even the best such models will not be able to predict specific instances of death but can only provide severity-adjusted expected death rates among groups of patients.

Predicting Who Dies Depends on How Severity is Measured: Implications for evaluating patient outcomes.

Annals of Internal Medicine, November 15, 1995, Vol. 123, No. 10: 763 - 770.

Iezzoni, Schwartz, Ash, Daley

*Results:* MedisGroups and the physiology score gave 94.7% of patients similar rankings.

Disease staging, MedisGroups, and the physiology score gave only 78% of patients similar rankings. MedisGroups and APR-DRGs gave 80% of patients similar rankings.

*Conclusions:* Some pairs of severity measures assign very different severity levels to more than 20% of patients. Evaluations of patient outcomes need to be sensitive to the severity measures used for risk adjustment.

(Page 764, paragraph 3)

Although the APR-DRGs measure was not initially developed to predict mortality, it is used for such analyses. For example, Iowa once required larger hospitals to produce MedisGroups data for severity-adjusted performance reports, but it switched in 1984 to using APR-DRGs—a less expensive, discharge abstract-based measure. This change was partially motivated by the perceived high cost of MedisGroups medical record reviews. Other states, such as Florida, also use APR-DRGs to evaluate health care provider performance.

(Page 767, paragraphs 4,5)

...Discharge abstract-based severity measures (disease staging and APR-DRGs) were slightly better able to predict death (measured by  $c$  and  $R^2$  values) than the clinical data-based measures (MedisGroup and the physiology score). In contrast, MedisGroups and the physiology score had better clinical credibility (relations with six clinical indicators of risk for death from AMI) than Disease Staging and the APR-DRGs. Thus, the discharge abstract-based measures had better predictive validity, and the clinical data-based measures had better clinical validity.

Differences between severity measures that occur on the basis of their data sources have important health policy implications, suggesting a trade-off between data costs and clinical credibility. Because discharge abstract data are routinely produced by hospitals, they are generally available, computer accessible, and inexpensive. These advantages have led some provider evaluation initiatives around the country (California, Connecticut, Florida, New Jersey, and Ohio) to use discharge abstracts. ... In addition, the clinical information contained in discharge abstracts is limited. Nonetheless, our finding that the discharge abstract-based measures were somewhat better able to predict death supports the choice of these measures.

#### The Risks of Risk Adjustment

JAMA, November 19, 1997, Vol 278, No. 19, Pages 1600 - 1607

Lisa Iezzoni, MD, Msc

*Results:* The severity measure called Disease Staging had the highest C statistic (which measures how well a severity measure discriminates between patients who lived and those who died) for acute myocardial infarction, 0.86; the measure called all Patient Refined Diagnosis Related Groups, had the highest for CABG, 0.83; and the measure MedisGroups, had the highest for pneumonia, 0.85 and stroke, 0.87. Different severity measures predicted different probabilities of death for many patients. Severity measures frequently disagreed about which hospitals had particularly low or high z scores.

Agreement in identifying low- and high-mortality hospitals between severity-adjusted and unadjusted death rates was often better than agreement between severity measures.

*Conclusions:* Severity does not explain differences in death rates across hospitals. Different severity measures frequently produce different impressions about relative hospital performance. Severity-adjusted mortality rates alone are unlikely to isolate quality differences across hospitals.

Pennsylvania's Declaration  
of Health Care Information  
A Commitment to  
Quality, Affordable,  
Health Care



## PENNSYLVANIA HEALTH CARE COST CONTAINMENT COUNCIL

Memo

ORIGINAL: 1995  
MIZNER  
COPIES: de Bien  
Harris  
Sandusky

To: Dave Wilderman  
Richard Dreyfuss  
Thomas Duzak  
Mary Ellen McMillen  
Richard Reif  
Donald Harrop, M.D.  
Melia Belonus

Fr: Leonard Borecki

Cc: Daniel R. Tunnell, PHC4 Chairman  
Clifford L. Jones, PHC4 Executive Director

Dt: 4-23-98

Re: Assessment of severity adjustment systems (Meeting at 9:30 AM, April 30<sup>th</sup>, Council office)

RECEIVED  
99 FEB 23 AM 11:55  
INDEPENDENT HEALTH CARE  
REVIEW COMMISSION

Thank you for agreeing to serve on the RFI committee which will make a final recommendation(s) to the Council at the May 7th meeting regarding its (the Council's) use of a system to adjust for patient differences in illness severity and/or risk. This is a very important responsibility and I appreciate your willingness to take time out of your busy schedules to be of assistance. As you know, we are meeting at 9:30 AM, Thursday, April 30th at the Council office.

Let me provide some background and how I think we should proceed on the 30th. I want to stress that what we have gone through is a Request for Information process, not a Request for Proposal. We have not solicited bids in order to select a new vendor; the Council's motion simply asks for an assessment of alternatives. After an extensive process, with thorough reviews by two independent and separate panels, we have narrowed the choice to two systems. One of these systems, MediQual, is the Council's current vendor whose system is based on data derived from patients' actual medical records (commonly known as a clinically based system). The second, 3M, is based on data derived from hospital billing records (commonly known as an administrative system).

### What has happened to date

1. This process began with the Council's motion of September 4, 1997 that stated:  
"to assess the costs, administrative burden and benefits of replacing the MediQual mandate with an alternative severity adjustment methodology that considers the cost burden to data suppliers and increases the relevancy, accuracy, timeliness, and usefulness of PHC4 data collection, analysis and reporting."

2. In November of last year, the Council published a Request for Information. Twenty companies submitted RFI responses (list of companies attached)
3. The Council convened a Severity Adjustment Assessment Panel (SAAP), composed of 13 members who were appointed by the PA Medical Society, the Hospital and Healthsystem Association of PA, the Hospital Council of Western PA, and the PA Health Information Management Association. The SAAP agreed to conduct the first reviews of the submissions and developed a review and scoring system. The staff was instructed by the SAAP to pre-screen the submissions for conformance with Commonwealth guidelines and the RFI requirements. They did this with the assistance of a SAAP panelist. Three proposals were eliminated. The SAAP then conducted their reviews of the remaining 17 submissions and in a meeting via conference call on March 6<sup>th</sup>, eliminated 11 submissions from the process and sent the six remaining submissions on to the Technical Advisory Group (TAG) for further consideration. Four of the six received unanimous consent (MediQual, Quadramed, International Severity Information Systems [ISIS] and 3M). Two were sent forward based on a 7-5 vote with one absent member (DynCorp and Michael Pine/Associates).
3. TAG then reviewed the remaining six submissions, focusing their attention on the four unanimous submissions. TAG met via conference call on April 15<sup>th</sup>. After extensive deliberation, the TAG recommended the following to the Council:

"The Technical Advisory Group (TAG) is suggesting to the Council that after reviewing the six RFI submissions, there is a classic distinction between administrative-based systems and clinically-based systems, that there are differences of opinion as to the value of each, and that TAG does not have a consensus of opinion on this. Therefore, the TAG is recommending that the Council continue to consider a clinically-based system and an administratively based system. We would note that of the clinical systems, MediQual was clearly the high performer, and of the administrative systems, the high performer was 3M. We are forwarding these two systems on to the Council for further review."

A summary of their discussion and TAG's advice to the Council is contained in TAG chairman Dr. David Nash's memo to Council chair Dan Tunnell (attached). The minutes of the meeting are also attached.

NINGO TUSCH NUISSETA

7. That brings us to the final step in this process leading up to the May 7<sup>th</sup> Council meeting. *Our committee is charged with picking up the ball from TAG and discussing whether 3M's system presents a viable alternative to the MediQual system.* The result of our discussion about this on April 30<sup>th</sup> will form the basis of our recommendation to the Council on May 7<sup>th</sup>.

#### Where we go from here

Let me frame the discussion we need to have.

Is 3M, an administrative-based system, an appropriate alternative to MediQual, a clinically-based system, that:

- 1) *meets the requirements of the Council's law (ACT 89) that defines its mission, approach and activities.*

The law requires the Council to collect data that permits analysis of:

- (a) **provider quality** – defined in the law as “the extent to which a provider renders care that, within the capabilities of modern medicine, obtains for patients medically acceptable health outcomes and prognoses, adjusted for patient severity, and treats patients compassionately and responsively.”
- (b) **provider service effectiveness** – defined in the law as “the effectiveness of services rendered by a provider, determined by measurement of the medical outcomes of patients - grouped by severity – receiving those services.”

Severity is defined in the law as “in any patient, the measurable degree of the potential for failure of one or more vital organs.”

So, point #1 is that the Council's selected severity adjustment system must enable the Council to meet its obligations under the law.

- 2) Will the Council be able to *publicly* report comparative quality of care-related data about hospitals and other appropriate health care facilities, physicians, and health plans with at least the same level of credibility and accuracy present in previous Council reports?
- 3) Does 3M's system provide such a greater cost savings to data sources, and such a reduced burden, that it warrants making a change, with the associated delays? These delays involve a Council Request for Proposal process, the promulgation of new regulations, education of hospital coding personnel, physicians, Council members and Council staff regarding a new system.

I have attached the following background information for your review including Dr. David Nash's letter as TAG chairman regarding TAG's recommendation, the TAG meeting minutes and RFI scores, list of the 20 RFI submitters, list of SAAP panelists, and the minutes from SAAP's final meeting.

These are the issues we need to consider. Thank you again for your willingness to participate. I look forward to our discussion on April 30th at 9:30 AM.

cc: Daniel R. Tunnell, Council Chairman  
Clifford L. Jones, Council Executive Director

777 East Park Drive  
P.O. Box 8820  
Harrisburg, PA 17105-8820  
Tel: 717-558-7750  
Fax: 717-558-7840  
E-Mail: STAT@PAMEDSOC.ORG



Pennsylvania  
MEDICAL SOCIETY®

150 years

*Legacy of Leadership  
Foundation for the Future.*

LEE H. MCCORMICK, MD  
*President*

April 22, 1998

JOHN W. LAWRENCE, MD  
*President Elect*

DONALD H. SMITH, MD  
*Vice President*

JAMES R. REGAN, MD  
*Chair*

ROBERT L. LASHER, MD  
*Secretary*

ROGER F. MECUM  
*Executive Vice President*

Mr. Marc P. Volavka  
Pennsylvania Health Care  
Cost Containment Council  
Suite 400, 225 Market Street  
Harrisburg, PA 17101

Dear Mark:

As per your telephone conversation with Dr. Harrop, I am confirming the Pennsylvania Medical Society's position, with regard to the evaluation of systems for collecting health care data.

The Society will only support a system that is capable of collecting clinically based and severity-adjusted data. As we have said on several occasions in the past, we are not now and have at no time in the past endorsed a particular vendor or system to fulfill these requirements. Our position remains based upon both the criteria noted above.

Should you have any questions, or need further information, please feel free to contact me or Bernie Lynch, in our Medical Economics Department.

Sincerely,

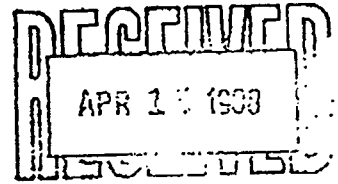
Roger F. Mecum  
Executive Vice President

cc: Donald E. Harrop, Chairman  
PMS Committee on Health Care Cost and Quality Data

Bernard Lynch, Associate Director  
Medical Economics




THE HOSPITAL & HEALTHSYSTEM ASSOCIATION OF PENNSYLVANIA



April 13, 1998

TO: TAG Member

FROM: Carolyn F. Scanlan,  President and Chief Executive Officer

SUBJECT: PHC4 RFI Evaluation

I am writing to you today to help clarify the position of The Hospital and Health System Association (HAP) regarding the PHC4 study of alternative severity adjustment systems. I understand that you are expected to complete your work by April 15, 1998. While I know you are well aware of the difficulty and cost incurred by hospitals in complying with the MediQual mandate, I wish to emphasize that we also share your concern regarding the validity and accuracy of a severity adjustment system for statewide public reporting. That concern is best summarized by the language adopted by the PHC4 Executive Committee on September 4, 1997, when it directed its staff to:

“Assess the cost, administrative burden, and benefits of replacing the MediQual mandate with an alternative severity adjustment methodology that considers cost burden to data suppliers and increases the relevancy, accuracy, timeliness, and usefulness of PHC4 data collection, analysis, and reporting.”

The six severity adjustment systems you received from the hospital severity adjustment advisory panel, include three clinically-based systems and three administratively-based systems. HAP prefers to see an arrangement whereby PHC4 utilizes an administratively based severity adjustment process which has been proven effective. Further, we would like to see that severity adjustment process centralized at PHC4. By so doing, hospitals would only be required to submit uniform billing records in a standard format to PHC4. If hospitals wanted their own data reports with the similar severity adjustment utilized by PHC4, they would have the ability to purchase their own user license. Hospitals that desire to continue using MediQual could continue to do so on their own. HAP's rationale is based on the following:

TAG Member  
April 13, 1998  
Page 2

- ▶ Discharge abstract-based measures have higher  $c$  and  $R^2$  scores than MedisGroups for both acute myocardial infarctions and coronary artery bypass graft. Thus, for two of the Council's public reports, discharge-based severity models have similar validity as clinically-based measures. (The Risks of Risk Adjustment, Lisa Iezzoni, MD, MSc, JAMA, November 19, 1997, Vol 278, No. 19, pp. 1604)
- ▶ There exists no single perfect severity adjustment method. Consequently, the literature demonstrates that comparisons of severity systems shows that for 80 percent of patients, MedisGroups and APR-DRGs gave similar rankings. (Predicting Who Dies Depends on How Severity is Measured: Implications for evaluating patient outcomes, Annals of Internal Medicine, November 15, 1995, Vol. 123, No. 10: 763 - 770. Iezzoni, Schwartz, Ash, Daley.)
- ▶ The literature shows that a potentially economical, and relatively powerful, predictor of in-hospital mortality for five conditions under study (stroke, lung cancer, pneumonia, acute myocardial infarction, congestive heart failure) could be achieved using an administrative data set augmented with 10 clinical variables. (Predicting In-Hospital Mortality: A comparison of severity measurement approaches, Medical Care, April 1992, Vol. 30, No. 4; pp. 357. Iezzoni, Ash, Coffman, Moskowitz )
- ▶ The PHC4 requirement for submitting administrative data to the Council in Harrisburg, and clinical abstract data to MediQual in Boston, is costly, cumbersome, and in light of major advances in severity adjustment methodologies over the past 13 years, outdated. Our survey of hospital CEOs indicates that when license fees, staffing, and technical costs are considered, average annual cost per hospital to comply with the mandate in its present form is approximately \$123,000. An administratively based system would speed up the data submission process for PHC4, and in turn, provide more timely data and reports for clinicians, academicians, and hospitals.

For your reference I have enclosed a copy of my March 3, 1998, letter to the PHC4 Executive Committee apprising them of HAP's position. I am also including a two-page synopsis of relevant sections of three articles referenced above. Copies of the three articles are also enclosed.



TAG Member  
April 13, 1998  
Page 3

Finally, we all want to improve the clinical outcomes for the patients of Pennsylvania. The selection of an administratively-based system will advance our mutual goal by allowing PHC4 to collect and report relevant information more quickly than before. As you participate in the RFI evaluation process, I would ask that you give serious consideration to an administratively-based system.

Enclosures

c: Daniel Tunnell  
Clifford Jones  
Marc Volavka

WESTERN PENNSYLVANIA

TAG RFI Submission Finalists Aggregate Scores

Sorted Alphabetically by Reviewer -- with average score

Reviewer	MediQual	3-M	ISIS	QuadraMed
J. Marvin Bentley, Ph.D.	85	67	83	63
David Campbell, M.D.	63	68	72	56
Paul Casale, M.D.	55	34	45	32
Donald Fetterolf, M.D., M.B.A.	99	78	96	93
James Grana, Ph.D.	45	60	40	60
George R. Green, M.D.	65	90	42	65
Sheryl Kelsey, Ph.D.	91	71	84	77
Judith Lave, Ph.D.	78	86	64	83
David B. Nash, M.D., M.B.A.	86	85	35	30
Average	74.11	71.0	62.33	62.11


Pennsylvania's Declaration  
of Health Care Information  
A Commitment to  
Quality, Affordable,  
Health Care



## PENNSYLVANIA HEALTH CARE COST CONTAINMENT COUNCIL

Memo

To: Daniel R. Tunnell  
Chairman  
PA Health Care Cost Containment Council

Fr: David B. Nash, M.D., M.B.A.   
Chairman  
PHC4's Technical Advisory Group (TAG)

Dt: 4-17-98

Re: TAG's assessment of severity adjustment systems

---

As chairman of the TAG, I first want to thank the Council for its confidence in entrusting the TAG with the important role of evaluating possible severity adjustment systems for future use by the Council.

I would also like to acknowledge the Severity Adjustment Assessment Panel (SAAP) for its hard work in reviewing the original 17 RFI submissions. Given the detail of these submissions, this was no mean feat.

I also need to thank my colleagues on the TAG for taking the time from their busy schedules to examine and score the six proposals forwarded to us by the SAAP.

The TAG met via conference call on April 15<sup>th</sup> to formulate its advice to the Council on the severity adjustment issue and I would like to summarize briefly the key points of the discussion. I then want to communicate TAG's recommendation regarding the Council's September 7th motion that in brief states "to assess severity adjustment alternatives to the MediQual system."

The Severity Adjustment Assessment Panel voted unanimously to send four Request for Information (RFI) submissions to the TAG: MediQual, 3M, International Severity Information Systems, and QuadraMed. The SAAP also voted 7-5 to send the submissions from Dyncorp and Michael Pine/Associates to TAG as well. All 9 TAG members reviewed the RFI material and rated them according to the same system as did SAAP. I must note that in my original memo to the TAG (attached), I asked that we concentrate our efforts on the four consensus submissions. As a result, Dyncorp and Michael Pine did not receive a score from all TAG members, although the four consensus proposals did. Eight of 9 TAG members were present on the conference call. (Dr. David Campbell was performing surgery at the time and could not join us, but sent his scores.)

Suite 400, 225 Marker Street Harrisburg, PA 17101

717-232-6787

www.phc4.org

FAX 717-232-3821

While TAG did spend some time discussing the merits of the individual submitters, the more important part of the discussion dealt in a broader sense with the pros and cons of a clinically-based system versus those of an administratively-based system. (Of the six submissions, three were clinical: MediQual, International Severity Information Systems, Dyncorp; and three were administrative: 3M, QuadraMed, Michael Pine/Associates.)

I think it is fair to say that the majority of opinion favored the scientific credibility of the clinical system, and conversely, the majority of opinion acknowledged the cost advantages to hospitals of the administrative approach. This must be viewed in the context that the measurement of severity is still an imperfect science, and that there is not a single perfect system.

The physicians of TAG, which include myself, Dr. Casale, and Dr. Fetterolf felt strongly that from our personal experience, administrative data systems have little credibility with the physician community. Previous public statements by TAG member Dr. David Campbell, Professor of Cardiothoracic Surgery at Penn State Geisinger Health System prompted Dr. Casale to state that Dr. Campbell shared that view as well, and I concur. I must say that Dr. Green does not join us in that view.

Others in the meeting felt that the predictive powers of each approach were not significantly different: therefore, the lower cost to hospitals of the administrative approach may outweigh the scientific and credibility advantages of a clinically-based system, and that the Council should consider this.

Dr. Kelsey pointed out that MediQual is the one system that has the flexibility to use both approaches.

TAG did not discuss the practical aspects of a change from MediQual to a different system, nor did we attempt to determine whether any of the other systems meet the precise requirements of the Act 89. These were not within our scope of expertise.

The final point I would make is that the question "*can an administrative-based system be used for public reporting in a CABG-like report, and would the TAG stand behind such a report?*" was pointedly raised but not addressed by the TAG members.

The TAG did endeavor to reach a consensus on a single approach. However, we were not able to reach an agreement that one approach was significantly better than the other. I have attached the scoring of each TAG member as well as the minutes of the discussion. I would note that the scoring shows MediQual as the top performer overall.

Therefore, the advice of TAG to the Council is the following:

The Technical Advisory Group (TAG) is suggesting to the Council that after reviewing the six RFI submissions, there is a classic distinction between administrative-based systems and clinically-based systems, that there are differences of opinion as to the value of each, and that TAG does not have a consensus of opinion on this. Therefore, the TAG is recommending that the Council continue to consider a clinically-based system and an administratively based system. We would note that of the clinical systems, MediQual was clearly the high performer, and of the administrative systems, the high performer was 3M. We are forwarding these two systems on to the Council for further review.

The TAG has left the Council with a choice to make and I believe the choice is between the following two principles: A decision in favor of the clinical system opts for scientific accuracy and credibility over cost issues, a decision favoring an administrative approach places a higher value on reducing the cost burden to hospitals over scientific accuracy and credibility, particularly with physicians.

The Council must decide which is more important in terms of its mission of public reporting as a means to reduce the cost and improve the quality of health care in Pennsylvania.

As I stated in the December 1997 TAG meeting, during a discussion over why the CABG report had been cancelled (a decision that I strongly opposed), I stated that the national view of experts in the field of clinical outcomes research and reporting is that the Pennsylvania Health Care Cost Containment Council represents the "gold standard" to all others across the country who are attempting to achieve what the Council has accomplished. My personal view is that the decision the Council makes should not compromise this reputation that we have all worked so hard for so long to build.

On behalf of my TAG colleagues, I personally want to thank the Council for the opportunity to participate in this very important process, the result of which will have, in my view, implications for years to come. Please call on me for any further assistance.

cc: TAG members

WESTON HOSPITAL

**Technical Advisory Group**

**Teleconference**

**April 15, 1998**

**Minutes**

**Participants:**

Via Telephone:

David B Nash, M.D., M.B.A.  
Paul N. Casale, M.D., F.A.C.C.  
Donald E. Fetterolf, M.D., M.B.A.  
James Grana, Ph.D.  
George Green, M.D.  
Sheryl F. Kelsey, Ph.D.  
Judith Lave, Ph.D.

PHC4 Conference Room:

J. Marvin Bentley, Ph.D.  
Marc P. Volavka  
Joseph Martin  
Susan Moore  
Flossie Wolf  
Paul McDowell

Via FAX:

David Campbell, M.D.

The conference call was convened at 10:00 a.m. Following introductions of the participants, Mr. Volavka turned the meeting over to Dr. Nash as Chair of the Technical Advisory Group (TAG). Dr. Campbell was unable to participate because he was performing emergency surgery. He faxed his scores to Mr. Volavka.

Dr. Nash thanked everyone for taking time to review and score the RFI proposals. The scores had been faxed to everyone so that they would be able to review them prior to the meeting.

Dr. Nash commented that in reviewing the scores the majority of the TAG members had ranked MediQual quite high. Dr. Green did not rate MediQual as high as the other proposals.

Dr. Nash commented that MediQual was rated highly on the clinical side. He stated that we have the richest clinical system available to satisfy the requirements under the law. The discussions were begun with this observation in mind.

Dr. Lave commented that if we must have a clinical system then there is no point in talking about an administrative database. Her concern is that when the two systems had gone head on, the clinical has not performed much better or much worse than systems that cost much less. The question is how does one evaluate the marginal improvement, if there is any. Dr. Lave noted that one expert, Dr. Lisa Iezzoni, suggests that there is not much additional value to a clinical system. Dr. Lave feels that there are several critical issues. One issue is that there is a major cost difference. The second issue is that the TAG received a letter from HAP, which seems to indicate that the hospitals are willing to be evaluated based on administrative data. Third we do not have any testimony from the hospitals in Pennsylvania that, in fact, stated that they have used the MediQual system as their quality basis. They are incurring a huge cost and not using the

information. She feels it is important to find out whether these data are useful to them for other purposes.

Dr. Nash commented that there are some obvious issues, 1) all hospitals already have MediQual in place so we need to consider the difference of taking out the system they already have and implementing the administrative dataset, if there is agreement that one is not superior to the other in analytic capabilities. Dr. Lave commented that there is a huge difference because the administrative system is less costly. Dr. Nash is concerned that the administrative dataset does not have the richness from a clinician perspective. Dr. Kelsey stated that one reason she rated MediQual the highest is that they did offer the option of both administrative and clinical databases. Her preference would be to go with the clinical data now but five years from now the administrative dataset may be richer as the science evolves. MediQual would have the ability to do both. There is no question that if we go with the clinical system, MediQual is clearly superior, according to Dr. Kelsey.

Mr. Volavka commented that we now have a version of MediQual's administrative system in place for calendar year 1998 for five MDCs with low or no risk of mortality. Hospitals already have the option not to abstract the clinical data but to use the MediQual Grouper, which is the administrative version. Council has already adopted some portion of that in recognition of one of the concerns that HAP had raised. Dr. Lave has a problem with this because the administrative data is based on charges and she does not believe charge data is meaningful data. Dr. Grana commented that Aetna US Healthcare had done an analysis a few years ago to determine if they could predict length-of-stay and mortality rates using the methodology similar to MediQual. They then introduced the MediQual score from PHC4 data. In the regression model, the MediQual score added no additional explanatory information to their model. He feels that administrative data, based on this experience, is capturing almost all of the necessary adjusters. He agrees that there is a need for a strong clinical element. Dr. Nash stated the issue of whether clinical databases are better than administrative databases is a difficult one to resolve.

Dr. Fetterolf commented that some of the administrative databases and their methods for analysis are very good but when you present to physicians and professional organizations, you are shot down almost immediately if you do not show some direct clinical relevance. The issue is the concern that adding clinical data elements would significantly drive up the cost, beyond the benefits derived from using administrative only databases. He disagrees with Dr. Lave about the cost to hospitals being so burdensome. He noted that the real cost per discharge is between \$10 and \$25. This is relative to what a hospital charges for a CBC. He believes we need some administrative data but in terms in real cost, it does not add that much. Dr. Grana stated that you could get a great deal of clinical relevance out of administrative data. Dr. Nash stated that if you talk to physicians and tell them it is administrative data, that is the end of the conversation.

Dr. Green commented that he was surprised the way his scoring came out. Originally, he felt that he would lean to the clinical side. He has become discouraged with the baggage that comes with the clinical system. He feels it is not current and at times does not make sense. He does not have confidence in the system. He feels it is very dependent on someone's clinical judgement. There has been a tremendous improvement in clinical charts in the past ten years. Dr. Green feels timeliness is the most important issue. The administrative system gives a lot more data more quickly for a lot less money.

Dr. Fetterolf feels there are a significant number of negatives with the MediQual system. In his opinion, the reason for this RFI is that the MediQual system has been problematic. These negatives need to be taken into consideration.

Dr. Nash suggested making a recommendation for both an administrative data set and a clinical data set.

Mr. Volavka, in contradiction to Dr. Lave's earlier contention that hospitals don't use the MediQual data, brought to the attention of the TAG members the survey that PHC4 conducted with the hospitals. The survey had an 84% response rate. One of the survey questions asked if the hospitals used the MediQual system to satisfy only the PHC4 mandate. Fifty-five hospitals responded yes and 97 responded no. Mr. Volavka will send a copy of the survey to the TAG members

Mr. Volavka responded to Dr. Grana's earlier comments about administrative systems and adverse events. He asked how an administrative system would define adverse events. Dr. Grana responded that clinical logic was introduced into these models so that certain things would count as adverse events. To capture these data, they used secondary diagnosis data. Mr. Volavka commented that PHC4 has looked in detail at our administrative data over the past four or five years and find that these data vary greatly across facilities and across the state in terms of exactly those kinds of codes. Some hospitals do not show any data and others show huge numbers. This is a major concern of the Council. Dr. Grana recognized this issue and noted that the Council would have to educate hospital personnel about correct coding. Dr. Bentley commented that these are some of the reasons the physicians question an administrative database.

Dr. Casale commented that the physicians have a concern with the administrative data set and really want the clinical data. He further stated that based on conversations with Dr. David Campbell that he believes Dr. Campbell concurred with this view.

Dr. Nash recognized that these are issues that all of the TAG members have been dealing with for years and may not be able to resolve them with consensus.

Following this lengthy discussion it was suggested that the TAG give the Council a recommendation on a clinical data set and an administrative data set.

A review of the scores indicates that on the clinical side, MediQual is clearly the favorite choice. On the administrative side, it is closer. We may need more cost information for QuadraMed and 3-M.

Dr. Grana agreed with the decision on MediQual for the clinical data set. He has a concern with 3-M's method for the administrative data set. He felt QuadraMed was more thorough in their explanation of how firms are doing and their acceptance nationally.

Dr. Bentley felt QuadraMed was mostly PR. He does not have as much confidence in their system. His experience with some of their reports has not been very positive.

Dr. Lave favors 3-M because she is familiar with their work and they do a lot of evaluation of their work. They also have experience doing publications.

Dr. Fetterolf, Dr. Lave, Dr. Kelsey, Dr. Bentley endorsed 3-M as the administrative data set recommendation. Dr. Grana felt we should get more information to find out whether they empirically derive their coefficients or whether they were subjectively derived. Mr. Volavka said we would not have time to follow-up with this suggestion due to the time constraints for our submission of the TAG recommendation to the Council.



Mr. Volavka asked, if the Council concluded that there is support for an administrative based system, is the TAG ready to say they would endorse a CABG like report, reporting outcomes for hospitals, surgeons and for plans based upon an administrative database system.

Dr. Green would want to have sufficient input into the way these are used. Dr. Fetterolf commented that no matter which one is chosen it won't be nearly as much use as it could be if we continue on in the way we do the studies and reports at present. He feels part of the recommendation should be to make sure that HC4 staff gets the report generating engines so that they can generate the kinds of reports that are suggested in the proposals rather than use these data to generate internal reports. Dr. Lave concurred with Dr. Fetterolf suggestion. Dr. Nash agreed.

Therefore, the final recommendation from TAG to the Council is that after reviewing the six RFI submissions, there is a classic distinction between administrative-based systems and clinically-based systems, that there are differences of opinion as to the value of each, and that TAG does not have a consensus of opinion on this. Therefore, the TAG is recommending that the Council continue to consider a clinically-based system and an administratively-based system. We would note that of the clinical systems, MediQual was clearly the high performer, and of the administrative systems, the high performer was 3M. We are forwarding these two systems on to the Council for further review.

Dr. Nash wanted to communicate to the Council that in part, their split vote reflects TAG's deep concerns regarding the past performance and cost of the MediQual system.

Staff will send the TAG recommendation to the Chair of the Council.

Dr. Nash and Mr. Volavka thanked everyone for their participation.

Respectfully submitted by:  
Roberta Six and Joe Martin  
April 20, 1998

# Pennsylvania

## Health Care Cost Containment Council

### Request for Information

PHC4 recently posted a Request for Information (RFI) to obtain information on risk adjustment/severity adjustment systems that would allow the Council to analyze the performance, reliability, operational cost, and financial viability of these systems with regard to meeting the Council's requirements. The deadline for responses was January 15, 1998.

At the close of the deadline, PHC4 received twenty responses from the following companies:

- |  |                      |
|--|----------------------|
| • 3M Health Information Systems                              | Alpharetta, GA       |
| • APACHE Medical Systems, Inc.                               | McLean, VA           |
| • Children's National Medical Center                         | Washington, DC       |
| • DynCorp  | Reston, VA           |
| • Greater New York Hospital Association                      | New York, NY         |
| • HBS International, Inc.                                    | Bellevue, WA         |
| • Health Care Data, Inc.                                     | Encinitas, CA        |
| • Health Data Research, Inc.                                 | Portland, OR         |
| • Health Systems Consultants, Inc.                           | New Haven, CT        |
| • Iameter, Inc.  | Atlanta, GA          |
| • IHPHSR   | Cincinnati, OH       |
| • International Severity Information Systems, Inc. (ISIS)    | Salt Lake City, UT   |
| • MediQual Systems, Inc.                                     | Westborough, MA      |
| • Medirisk   | Chicago, IL          |
| • Michael Pine Associates, Inc.                              | Chicago, IL          |
| • QuadraMed Corporation                                      | Neptune, NJ          |
| • The Joint Commission on Accreditation of Health Care Org's | Oakbrook Terrace, IL |
| • The Medstat Group  | Nashville, TN        |
| • The Society for Thoracic Surgeons                          | Chicago, IL          |
| • Uniform Data System for Med. Rehab./JDSmr                  | Buffalo, NY          |

PHC4 remains committed to the review of the responses to this RFI and are taking every precaution to ensure a fair process for everyone involved. While the specific details about the process have not been finalized at this time, we will keep you informed of all future actions. In the meantime, if you have questions, please call or e-mail either Joe Martin or Marc Volavka.

Robert Corrato, MD  
Thomas Jefferson University  
1015 Walnut Street  
Philadelphia, PA 191075099  
T 215-955-0240 F 215-923-7583

Mr. Dan Garofalo  
Hospital & Healthsystem of Pennsylvania  
4750 Lindle Rd., POBox 8600  
Harrisburg, PA 171058600  
T 717-561-5307 F 717-561-5216

Ms. Susan L. Lawrence  
Lehigh Valley Hospital  
Cedar Crest & I-78, POBox 689  
Allentown, PA 181051556  
T 610-402-1765 F 610-402-8613

Mr. Bernard Lynch  
Associate Director, Medical Economics  
Pennsylvania Medical Society  
777 East Park Drive, POBox 8820  
Harrisburg, PA 171058820  
T 717-558-7750 F 717-558-7841

Ms. Joan K. Richards  
President  
Crozer-Chester Medical Center  
One Medical Center Boulevard  
Upland, PA 190133995  
T 610-447-2785 F 610-447-2234

Carl Sirio, MD  
Dept. of Anesthesiology & Critical Care  
University of Pittsburgh School of Medicine  
DeSoto & O'Hare Streets  
Pittsburgh, PA 15213  
T 412-647-0112 F 412-647-8060

Ms. Nina Zimmer  
Thomas Jefferson University Hospital  
# 2130 Gibbon  
111 S. 11th Street  
Philadelphia, PA 19107  
T 215-955-5486 F 215-923-9270

Ms. Mary Anne Darragh, RRA  
Senior Vice President  
AHERF  
120 Fifth Ave., Suite 2900  
Pittsburgh, PA 15222  
T (412) 359-6876 F 412-359-3444

Ms. Joanne Klinedinst  
Mgr, Application, Development & Support  
Doylestown Hospital  
595 West State Street  
Doylestown, PA 18940  
T 215-345-2138 F 215-345-2040

Mark Lyles, MD  
Hospital of the University of PA  
# 1, Founders  
3400 Spruce Street  
Philadelphia, PA 19104  
T 215-662-3998 F 215-349-5864

Mr. Robert Morrison  
Data Analyst  
Centre Community Hospital  
1800 East Park Avenue  
State College, PA 16803  
T 814-231-3189 F 814-231-7077

Joan Silver, RN  
Director for Outcomes Management  
Pinnacle Health Systems  
P.O. Box 8700  
Harrisburg, PA 171058700  
T 717-231-8722 F 717-231-8768

Ms. Susan Vasco  
Saint Claire Hospital  
109 Locust Street  
Pittsburgh, PA 15241  
T 412-344-6600x1156 F 412-572-6584

**Meeting Notes and RFI Scores of the four sub-groups  
of the Severity Adjustment Assessment Panel  
(SAAP).**

Each sub-group rated 4 submissions plus MediQual.  
The SAAP decided that all panelists should review  
and rate MediQual's submission.

**Summary of All Severity Adjustment  
Sub-Panel Scores by Rank**

Average Score	Company Name
98.00	QuadraMed
91.50	3M
90.00	Iameter...
82.70	DynCorp
82.02	Mediqual Avg
78.00	Health Care Data
77.75	ISIS
73.50	Health Systems Consultants
70.00	Greater NY Hospital Assoc
69.75	Michael Pine Assoc
68.00	Medstat
66.00	Apache
58.00	HBS
43.05	JCAHO
29.65	Medirisk
n/a**	Children's Natl Medical Center
n/a**	UDSMR

Mediqual Average Score by Group	Group
86.27	B
86.00	C
74.50	A
81.3	D
82.02	Avg Total

WESTON HOSPITAL

\*\* RFI didn't meet application criteria

# Group A

## Meeting Notes

### Purpose of the Meeting

The Pennsylvania Health Care Cost Containment Council issued a Severity Adjustment Request for Information (RFI) in response to the Council's September 4, 1997 motion: "to assess the costs, administrative burden and benefits of replacing the Mediquel mandate with an alternative severity adjustment methodology that considers the costs burden to data suppliers and increases the relevancy, accuracy, timeliness, and usefulness of PHC4 data collection, analysis and reporting."

The specific purpose of this meeting was to assign a numerical rate to five Severity Adjustment proposals, one being Mediquel and four alternates.

### Meeting Summary

Joe Martin reviewed the Severity Adjustment Sub-panel Scores with the participants. Each member evaluated the RFI's from the Children's National Medical Center; Greater New York Hospital Association; Health Systems Consultants; 3M; and Mediquel. 3M scored the highest at 91.5 with Mediquel coming in second at 74.5.

### Severity Adjustment Sub-panel Scores - A

	Children's Nat Med Ctr	Greater NY Hosp Assoc	Health Sys Consultants	3M	Mediquel
Carl Sirio, MD	44	64	78	90	87
Dan Garofolo	0	76	63	93	67
Joan K. Richards	69	85	63	88	64
Robert Corrato, MD	0	55	90	95	80
Average Score		70	73.5	91.5	74.5

Mr. Garofolo and Dr. Corrato did not rate the Children's Medical Center because they felt the Center didn't meet the application requirements. Mr. Garofolo commented that he felt it was a separate severity system for a different patient cost that wouldn't reduce administrative burden or hospital costs. Dr. Sirio agreed that the Center was too specialized but stressed that the Council shouldn't discard the notion of niche products.

Mr. Martin asked for comments on the point values. Dr. Sirio noted that from a methodological standpoint he felt it would be in the council's best interest to review the two DRG systems, Health System Consultants and 3M, together. Ms. Richards expressed concern about using a

## Group A

system that requires special abstractions and doesn't use already existing systems, which may be the case with Health System Consultants. Dr. Sirio agreed that the details needed to be worked out at a later time but was recommending that the Council look at both systems.

Joe Martin explained that this teleconference was simply for the panel to come to agreement on the average point values of the submissions. A full panel conference call will be held at a later date to review the four different panel's submissions and discuss which submissions would go forward to tag. All participants agreed that they were comfortable with the averages and these scores should be announced to the full panel. Mr. Martin noted that this group scored Mediqua somewhat lower than the other two panels that have met thus far. However, Mediqua has scored first or second across the board.

Mr. Martin thanked the participants for the significant amount of time they spent reviewing the RFI's.

*Meeting Notes Prepared by Staff Member:*

*Susan Moore  
March 2, 1998*

PRINTED REVERSE SIDE

Group B

General Comments

Nina suggested that the scores by category be put together and averaged out. The above information is the result of this request.

Sue suggested that if we need to do a similar process in the future, we should not allow quite as much information to be submitted. It was however noted that the much of the information was required to make a reasonable decision. She pointed out that Medirisk and JCAHO were not appropriate systems to replace mediquial because they admittedly do not meet the mandate.

The final findings from our group about the processes that we scored were that Mediquial is the most comprehensive coverage for severity scores, but if the council is simply looking for something inexpensive that would only satisfy the mandate, Michael Pine Associates would be an alternative to consider. The final ranking of the systems were:

- 1. Mediquial
- 2. ISIS
- 3. Michael Pine
- 4. JCAHO
- 5. Medirisk

Severity-Adjustment-Sub-panel Scores - B

	ISIS	Medirisk	Michael Pine Assoc.	JCAHO	Mediquial
Nina Zimmer	89	27	62	32	88
Robert Morrison	81	41.6	73	60.2	90.1
Susan Lawrence	70	17	74	30	86
Lori Miller	71	33	70	50	81
Average Score	77.75	29.65	69.75	43.05	86.275



Group C

**Summary of Severity Adjustment RFI  
Conference Call**

**Participants:**

Wendy Whitmer-PHC4 Staff  
Joe Martin-PHC4-Staff  
Bernard Lynch-PMS  
Susan Vasco-St. Clair Hsp

The conference call was held on Friday, February 20 at 10:00. One of the participants of our group, Joanne Klinedinst of Doylestown Hsp, was unable to participate in the conference call because she had not yet completed her evaluations. It was determined that the conference call would continue without her and she would fax her completed evaluation to us with-in the week.

Joe began the call with a thank you to all of the participants for their time and effort involved with this task. Wendy then gave them the average scores that were received and rated the vendors in the order of their scores. She then asked them if they had questions regarding with the ratings. All of the participants were in agreement because the top two vendors were the same for all members. Because we were missing Joanne's evaluations, Wendy informed the group that when she received those scores she would average them in and if the rating order changed she would contact them. When Wendy received the ratings from Joanne, the order of the vendors scores did not change so the members of the group were not recontacted.

At the end of the call, Joe asked the members to send in their comments concerning their individual evaluations to Wendy and he again thanked everyone for their time.

Attached to this summary is the summary of the scores and copies of the completed evaluations that were received.

Group C

### Severity Adjustment Sub-panel Scores - C

	MEDSTAT	APACHE	HBS	Health Care Data	Mediqual
Bernard Lynch	60	55	65	85	90
Joanne Klinedinst	80	76	76	75	83
Susan Vasco	80	86	54	81	87
Wendy Whitmer	51	48	36	70	85
<b>Average Scores</b>	<b>68</b>	<b>66</b>	<b>58</b>	<b>78</b>	<b>86</b>

### Ranking Order Of Vendors

Mediqual 86 points  
Health Care Data 78 points  
MEDSTAT 68 points  
APACHE 66 points  
HBS 58 points

WENDY WHITMER

Group D

## Severity Adjustment Sub-Panel Conference Call Minutes

March 2, 1998 2pm

### Attendees:

Kim Murawski  
Joan Silver  
Mark Lyles  
Bonnie Tatenal  
Mary Anne Darragh  
Terri Yencha  
Joe Martin

### Purpose:

The purpose of this conference call was to discuss the ratings of our five RFI responses, including DynCorp, Iameter, QuadraMed, Uniform Data System for Medical Rehabilitation (UDSMR), as well as MediQual.

It was the intention of this group to use this opportunity to select the RFI response(s) that warranted a further look by the other members of the panel and discard the responses that we felt did not provide an equal or improved option to the current system.

### Minutes:

Joan Silver and Kim Murawski personally joined Joe Martin and Terri Yencha at PHC4 for the conference call.

We began the conference by discussing that since the UDSMR system was too niche-specific as a rehab-only system, it should be discarded at this time. DynCorp was also discussed as being too much of a focus project, and that the photocopying and scanning involved was too in-depth and should also be discarded.

We all agreed that QuadraMed received the highest ranking among each of us. Joan liked their honesty in revealing their data problems and the modifications they have performed in the past to correct them, whereas all the other responses commented that they had no data problems.

Mary Anne agreed that QuadraMed seemed responsive to the user community and appeared to be the best balance of a good system with the cost of the product. Furthermore, it appears to be a growing company with an eye on the future. Mary Anne commented that it would probably be of our best interest to discuss this company with the New Jersey community in order to get their perspective.

Bonnie added that she heard the QuadraMed system discourages coding of complications, however it was not clear whether it was the users who were not coding the complications or if the system simply adjusts to identify these complications.

Group D

*Further action should be taken to answer these coding questions regarding QuadraMed, possibly with someone from the New Jersey hospital community.*

Iameter was also discussed and ranked second by most of the panelists. We felt it was also in contention as a good system to be examined further. Only one panelist had any experience with this system and new measures have been included since that time.

MediQual was ranked third by most of the panelists (with the exception of Terri who ranked it second mainly because she does not use this system in her daily work functions and only rated the system on the RFI response).

Terri announced that the next conference call will take place on Friday, March 6 at 1:30. The minutes and results of the other panelist will be faxed prior to the conference call in order to discuss the submissions that require a further look by the other members of the panel.

**Conclusion:**

With the conclusion of the discussions, we feel that the responses from our submissions that warrant a further look by the other members of the panel include QuadraMed, Iameter, and MediQual.

WELTON HOSPITAL ADMIN

Group D

### Severity Adjustment Sub-Panel Scores - D

	QuadraMed	Iameter	MediQual	DynCorp	UDSMR
Kim Murawski	102	95		83	
Joan Silver	99	91	88	80	
Mark Lyles (and Bonnie Tatenal)	90	88	65		
Mary Anne Darragh					
Terri Yench	101	86	91	85	41
Average Score	98	90	81.3	82.7	N/a

# 1995

# Severity-Adjustment of Hospital Outcomes: Do APR-DRGs and MedisGroups Yield Similar Results?

Study conducted by  
**the Colorado Hospital Association**

**Bonnie B. McCafferty, M.D., M.S.P.H.**  
Principal Investigator

and

**W. Michael Boyson, M.H.A.**  
Director of Data Services and Research

**Richard N. Camfield**  
Programmer/Analyst

ORIGINAL: 1995  
MIZNER  
COPIES OF COVER: de Bien, Harris  
Sandusky.  
Original in file

INDIRECT  
REVIEW/COMMISSION

99 FEB 23 AM 11:55

RECEIVED

October 1995

February 1992  
Volume 18/Number 2

RECEIVED

99 FEB 23 AM 11:55

INDEPENDENT REGULATORY  
REVIEW COMMISSION

Hospital Assoc. of PA  
Library

# QRB

Quality Review Bulletin  
Journal of Quality Improvement

ORIGINAL: 1995

MIZNER

COPIES: de Bien  
Harris  
Sandusky

---

## ARTICLES

---

A Description and Clinical Assessment of the Computerized Severity Index™

---

The Relationship Between Reported Problems and Patient Summary Evaluations of

---

Hospital Care

---

Self-Reported Versus Actual Test Ordering Behavior Among Primary Care Clinicians

---

Seeking Consensus on Important Aspects of Nursing Care

---

## DEPARTMENTS

---

### Letters

---

Meeting Update: National Demonstration Project and Joint Commission Forums Celebrate

---

Successes and Address Future Needs in Quality Improvement, I: National Demonstration

---

Project's Third Annual National Forum on Quality Improvement in Health Care

---

### Abstracts

---

### Meetings Calendar

---

---

---

---

## A Description and Clinical Assessment of the Computerized Severity Index™

*Prior to the implementation of Medicare's Prospective Payment System, methods for measuring severity of illness received little attention in the health care industry. Since the early 1980s, however, severity measurement has been a topic of broad concern to hospitals, payers, purchasers and policy makers. For several years, there was an expectation that the Health Care Financing Administration (HCFA) would choose one of the then available severity measurement systems for adjusting within diagnosis-related group hospital payments, and competition among system vendors for this designation became known as the "severity horse race." In the latter half of the decade, as it became apparent that HCFA would not select a commercial*

Lisa I. Iezzoni, MD, MS, is Co-Director, Division of General Medicine and Primary Care, Beth Israel Hospital, Boston, Massachusetts; Assistant Professor, Department of Medicine, Harvard Medical School, Boston; and a member of the QRB Editorial Advisory Board. Jennifer Daley, MD, is Assistant Professor, Division of General Medicine and Primary Care, Beth Israel Hospital; Assistant Professor, Department of Medicine, Harvard Medical School; and Assistant Professor, West Roxbury Veterans Affairs Medical Center, West Roxbury, Massachusetts. Please address requests for reprints to Dr Iezzoni, Department of Medicine, Beth Israel Hospital, 330 Brookline Avenue, Boston, MA 02215.

This research was supported by the Health Care Financing Administration, Office of Research, No. 18-C-99526/1-05. The views expressed are solely those of the authors.

The authors thank Susan D. Horn, PhD, June Buckle, ScD, RN, and Richard Averill, MD, for assisting in their understanding of the Computerized Severity Index™, and Mark A. Moskowski, MD, for his comments on the development of the manuscript.

*severity system for hospital payment, the focus shifted toward severity systems' use in hospital performance assessment. Purchasers and health care coalitions initiated "buy-right" programs using severity-adjusted outcome data, state data commissions began incorporating severity measures in their public data bases, and commercial vendors of health care data began offering evaluations of hospital quality based on severity-adjusted mortality, readmissions, and other outcomes. By 1991 at least two large payers, the Blue Cross plans of Minnesota and Michigan, had established programs to base hospital payment rates partially on assessments of severity-adjusted outcomes, financially rewarding facilities that report fewer than expected adverse outcomes and penalizing those whose outcomes are worse than expected.*

*As severity systems gain acceptance among payers and purchasers, as well as among hospitals interested in internal monitoring of physician practice efficiency and in continuous quality improvement programs, studies like the one presented here become particularly important. Severity is not a well-defined concept, and its measurement is conceived and operationalized in very different ways by different system vendors. Potential users of severity data should recognize the limitations, as well as the strengths, of individual measurement systems being considered. In their description of the Computerized Severity Index™, Iezzoni and Daley contribute to our understanding of one of these systems. — J. William Thomas, PhD, Department of Health Services Management and Policy, School of Public Health, The University of Michigan, Ann Arbor.*

While most recent policy initiatives are designed to constrain the health care delivery system, growth of health care data networks has been encouraged.<sup>1</sup> The centerpiece of this expansion lies in severity of illness information, which has three major goals:<sup>2-7</sup> (1) to improve the fairness of payment to hospitals that are treating patients who are sicker, and thus generally more expensive to care for, within a given diagnosis-related group (DRG); (2) to adjust for the risk of poor outcomes (for example, death); and (3) to indicate undesirable outcomes, such as morbidity or increases in severity during hospitalization. Payers particularly hope that severity of illness information can help identify high-quality providers.<sup>4-5</sup>

Spurred by this interest, a number of severity measurement techniques have been developed.<sup>4-17</sup> One system that should soon gain prominence is the Computerized Severity Index™ (CSI).<sup>14-23</sup> The purpose of this article is to describe the CSI, to review critically various aspects of its methodology, and to suggest areas in which further study and refinement are needed. (The description and critique of the CSI relate to the version in use in July 1989.)

### Background

In 1982 Susan D. Horn, PhD, and colleagues at the Johns Hopkins Medical Institutions (Baltimore) began to develop the CSI system. The impetus for its creation was the desire to respond constructively to criticisms of their earlier (1980s) system—the manual Severity of Illness Index (SOII).<sup>24-27</sup> Intended to link severity and resource use, the SOII rated severity independent of diagnosis along seven dimen-



sions, including stage and complications of the principal diagnosis, dependency on nursing services, and rate of response to therapy. The SOII was criticized for the subjectivity of the required judgments, its dependence on treatment variables and patient outcome, and its inability to distinguish the consequences of poor quality care.<sup>29</sup>

The CSI was designed not only to transcend these criticisms but also to respond to pressing federal and local health policy concerns. Its development commenced during the fractious period when Medicare proposed and adopted DRG-based payment schedules, with the attendant public apprehension about potential quality shortfalls and inadequate payment to hospitals with caseloads of severely ill patients. The CSI's developers explicitly wanted the CSI to be capable of refining DRG-based reimbursement and assisting in widespread quality assessment. This interest in resource needs and attributes of patient risk is reflected in the following definition of "severity" chosen for the CSI: *The treatment difficulty presented to physicians due to the extent and interactions of a patient's diseases.*

This conceptual definition was operationalized during the CSI development by focusing on length of stay as a proxy for treatment difficulty—that is, clinical factors associated with higher lengths of stay were assigned higher severity levels.

There were several underlying principles that guided the CSI's development, given its diverse, policy-oriented objectives (see Table 1, p 46). Most important was the conviction that severity is a diagnosis-specific concept. The CSI's developers believed that different clinical parameters are deranged in different diseases, and that the same derangement may have a varying clinical impact in different diseases (for example, a temperature of 38.8° C (102° F) is a very severe finding in leukemia, but moderately severe in pneumonia). Therefore the CSI was to rate individually each of a patient's diseases—as well as provide an overall assessment of severity. This differentiates the CSI from "generic" systems, which judge severity independent of diagnosis (such as

the Acute Physiology and Chronic Health Evaluation [APACHE]<sup>9-10</sup> and MedisGroups<sup>13-15</sup> systems).

#### CSI Development

In the mid-1980s the initial steps in the CSI's development began and involved review by research nurses at Johns Hopkins of the literature and standard medical textbooks to identify factors indicative of severity in each disease. Physicians reviewed these factors and specified their relationships to severity. The initial scoring approach—developed using purely clinical judgment—was tested on approximately 15,000 cases from five university hospitals.<sup>22</sup> In 1988 further refinement of the CSI algorithm was performed using data collected on approximately 74,000 cases from 25 hospitals in New Jersey.<sup>23</sup>

The weights assigned to specific clinical factors were calibrated by examining their relationships to length of hospital stay, with the goal of assigning higher severity scores to cases with longer stays. In 1985 Health Systems International\* (New Haven, Connecticut) began to input the complicated clinical logic of the CSI into a software program that can be processed on a personal computer.

#### Description of the CSI

The CSI methodology can be distilled into the following five components:

1. A way to define diagnosis. The developers of the CSI chose the *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) as its diagnostic lexicon. The CSI combines the more-than 10,300 ICD-9-CM codes into more than 820 disease groups, which serve as the CSI's diagnostic units.

2. A method to develop a list of severity criteria for each disease group. Criteria are arrayed in the form of a severity matrix for each disease group,

\* Recently renamed 3M/Health Information Systems (HIS) and relocated to Wallingford, Connecticut, 3M/HIS is responsible for the technical design of the CSI software (which works on a Xenix operating system), marketing, and client support. Dr Horn retains primary responsibility for clinical refinement of the CSI as well as the training of abstractors.

which lists both the specific clinical criteria and the severity level (1,2,3, or 4) ascribed to each criterion (see Table 2, p 46).

3. A method for computing disease-specific severity. This simple algorithm—based on the criteria identified for each disease group and their associated severity levels—assigns a score from 1 to 4 to each ICD-9-CM code that is listed in a patient's discharge abstract.

4. A method for computing the overall patient severity. This is a complex computerized algorithm based on the number and types of the patient's diseases and their disease-specific severity levels. Much of the specific weighting scheme derives from the principal diagnosis (scores range from 1 to 4).

5. A way for examining severity over the course of the hospitalization. The three major approaches are admission severity, maximum severity (based on the worst physiologic or other types of derangements, regardless of when they occur), and discharge severity. (Various aspects of the CSI are described in greater detail below.)

Since the severity matrices narrowly focus on the specific manifestations of the associated disease, the CSI depends on complete diagnostic coding. All the varied aspects of diseases, especially complications outside the organ system in which the disease originates, must have their own ICD-9-CM codes. For example, the severity matrix for lung cancer contains neither information about distant metastases nor their manifestations; separate diagnostic codes are required for these complications. Despite this, the CSI attempts to compensate for ICD-9-CM redundancy (for example, codes are available for pathophysiology, anatomy, clinical diagnoses, signs and symptoms, and physical examination findings<sup>23,20</sup>). In producing the overall severity score, the CSI does not permit use of multiple codes from the same organ system or related disease group.

In addition, the CSI attempts to transcend the pitfalls of inaccurate ICD-9-CM coding that have been the frequent subject of study and comment—especially with the advent of DRG-

**Table 1. Underlying Principles of the Computerized Severity Index™**

<ul style="list-style-type: none"> <li>• Severity is a disease-specific construct.</li> <li style="padding-left: 20px;">Severity levels within diseases are not necessarily comparable across diseases.</li> <li>• Different diagnoses should receive different weights in rating overall patient severity.</li> <li>• Overall patient severity should relate to the number of organ systems involved: a severe problem in a single organ system should generally receive less weight than problems in multiple organ systems.</li> <li>• Disease-specific severity should relate to the number of different manifestations of the disease: a single severe derangement should not dictate the severity rating.</li> <li>• Severity should not depend on treatment or major procedures.</li> <li>• The overall severity score must be simple, and comparable across broad groups of patients (for example, adult medical and/or surgical patients).</li> <li>• Severity measured at different points in time (for example, at admission, at discharge) is useful for different policy related and other purposes.</li> </ul>
--

**Table 2. The Meaning of the Computerized Severity Index™ Severity Levels\***

1 = Normal to mild
2 = Moderate
3 = Severe
4 = Life-threatening+
<p>*A 0 is assigned if the accuracy of the <i>International Classification of Diseases, Ninth Edition, Clinical Modification</i> (ICD-9-CM) diagnostic code is questioned.</p> <p>+The overall maximum and discharge severity for all patients who die in-hospital is automatically set to 4.</p>

based payment schedules.<sup>21-24</sup> If none of the questions asked on a given severity matrix are answered in the affirmative, this purportedly calls into question the accuracy of the ICD-9-CM diagnostic code, and the CSI automatically assigns the code a rating of 0. The 0 is not itself an indicator of severity but instead flags a probable coding error.

The most common source of ICD-9-CM codes is the discharge abstract, but discharge diagnoses may not be appropriate for rating severity on admission. For example, if a patient with an acute myocardial infarction (AMI) develops pneumonia one week after hospitalization, the admission severity should not consider the pneumonia. Since the ICD-9-CM codes are used to trigger the disease-specific CSI queries, it may be necessary to distinguish conditions present on admission from other discharge diagnoses.

As shown in Tables 3 (p 47) and 4 (p 3), the severity matrices are arranged as grids. Data are collected only for those variables contained in

the matrices associated with the patient's diagnoses. The individual data elements in the matrices are called *factors*, and related factors are grouped into *criteria* (for example, hemoglobin and hematocrit are factors that comprise one criterion). Criteria are unrelated to treatment and most are drawn from physical examination findings; signs, symptoms, and vital signs; and routine or relatively noninvasive testing. (A Help File [a glossary] defines the more descriptive terms.)

The ranges for continuous variables, which constitute different severity levels, may vary by disease group. For example, a hematocrit value of less than or equal to 15.0% is considered at severity level 4 for anemia, gastrointestinal hemorrhage, and breast cancer, while a hematocrit less than or equal to 20.0% is at level 4 for shock, thoracic aneurysm, and ectopic pregnancy. For some continuous variables, such as heart rate and arterial pH, two readings must be found to assign the factor to a severity level—so as to avoid excessive emphasis on a single

aberrant finding.

Severity ratings are based on criteria, regardless of how many factors are identified within a criterion. Disease-specific severity scores for each ICD-9-CM code present are computed by examining the severity levels of the criteria identified from the disease's severity matrix. Two criteria from a given severity level are generally required for the severity to be assigned to a diagnosis (for example, two level-3 criteria are required to assign that diagnosis to severity level 3). If only one criterion is identified at a higher level, the diagnosis is assigned to the next lowest severity level at which a criterion is found. (Refer to Table 2 for the meaning of the severity scores.) Although all disease-specific scores range from 1 through 4, the severity implied by these scores cannot necessarily be compared across diseases.

Patients are assigned "overall severity scores" based on the type and severity of their principal and secondary diagnoses. Overall scores also range from 1 through 4 (Table 2), and can be compared across broad classes of patients, such as adult medical and/or surgical patients. The technical algorithm for score computation is very complex and is described in detail elsewhere.<sup>21</sup> Briefly, a large portion of the algorithm is devoted to narrowing down the list of ICD-9-CM diagnoses to avoid redundancy and double counting. The major principle guiding overall score computation is that multiorgan system failures—as manifested by serious disease in two or more unrelated diagnoses—should generally be weighted more heavily than a single, severe disease. For example, if a patient has a single diagnosis of AMI rated as life threatening (severity level 4), the overall severity is 3. If a serious unrelated secondary diagnosis were present, the overall score could be increased to 4.

A second principle is that different diagnoses should receive different weights in computing the overall score. For example, in contrast to the scoring for AMI as described above, a single diagnosis of diabetes mellitus with coma at disease-specific severity level 4 results in an overall severity of 2. In

addition, the contribution of unrelated secondary diagnoses to overall severity varies depending on the principal diagnosis. For example, suppose the secondary diagnosis with the highest severity is pneumococcal pneumonia, at severity level 2. If the principal diagnosis is diabetes mellitus with renal manifestations, the pneumococcal pneumonia is ignored in computing overall severity. However, if the principal diagnosis is diabetes mellitus with coma, the pneumonia adds one point to overall severity.

A third principle is that each criterion should be used only once in computing the overall score. For example, a criterion such as hematocrit appears in many severity matrices. However, for a patient with multiple diagnoses that include hematocrit as a severity factor, the hematocrit level may be used only by one of these diagnoses in computing the overall severity. A complicated linear programming algorithm is employed to allocate such criteria to maximize overall severity.

Both disease specific and overall severity scores can be assigned at various points during hospitalization, that is, depending on their intended use. The three most common approaches are (1) admission severity, (2) maximum severity, and (3) discharge severity. Admission severity examines the first calendar day plus 24 hours, including the worst values over this period. Accurate conduction of this review necessitates designation of which ICD-9-CM codes pertain to this period, and the CSI software allows the reviewer to flag which of the discharge diagnoses were present upon admission. Patients who die during the admission period are assigned an overall admission score of 4.

Maximum severity encompasses the hospitalization period, including all discharge diagnoses. The maximum severity approach necessitates collection of the most deranged values from the entire stay, regardless of which day they occurred. For example, serum potassium valued could be drawn from day 1, arterial oxygenation from day 2, and hematocrit from day 3, if the most aberrant values occurred on these different days. Thus,

Table 3. The Computerized Severity Index™ Matrix for Lung Cancer

Malignant Neoplasm of Lung						
	162.2-164.0	164.2-164.9	165.8	194.6	195.1	197.0-197.2
	212.3-212.6	214.2	215.4	227.6	235.7-235.8	239.1
Category	1	2	3	4		
Cardiovascular	• JVD† ≤2cm	• JVD 3-5cm	• JVD 6-9cm	• JVD ≥10cm		
			• 1+-2+ edema of neck, face and upper extremities	• 3+-4+ edema of neck, face and upper extremities		
	• Pulse rate <100; ST segment changes on EKG	• Pulse rate 100-129; PACs, PAT, PVCs on EKG	• Pulse rate ≥130	• Life-threatening arrhythmias or hypotension		
Digestive	• Dysphagia NOS		• Unable to swallow solids	• Unable to swallow liquids		
Fever	• ≤100.4 and/or chills	• ≥100.5 and/or rigors				
General	• Weight loss ≤5.9%	• Weight loss 6.0%-15.9%; cachectic	• Weight loss 16.0%-20.9%	• Weight loss ≥21.0%		
Labs						
ABGS	• pH 7.35-7.45	• pH 7.46-7.50 7.34-7.25	• pH 7.51-7.60 7.24-7.10	• pH ≥7.61 ≤7.09; pO2 ≤50		
Chemistry	• Albumin ≥3.2g/dl	• Albumin 3.1-2.9g/dl	• Albumin 2.8-2.5g/dl	• Albumin <2.4g/dl		
Hematology	• HCT ≥30.0%; HGB ≥10.0g/dl	• HCT 29.9%-20.1%; HGB 9.9-6.6g/dl	• HCT 20.0%-15.1%; HGB 6.5-5.1g/dl	• HCT ≤15.0%; HGB <5.0g/dl		
Neurology	• Generalized weakness • Hoarseness					
Respiratory		• Dyspnea on exertion; stridor; decreased breath sounds <=50%/ <3 lobes	• Dyspnea at rest; decreased breath sounds >=50%/ >3 lobes	• Apnea; absent breath sounds >50%/ >3 lobes		
	• White, thin, mucoid sputum; productive cough	• Hemoptysis NOS	• Frank hemoptysis			

†JVD, jugular venous distention; ST, stress test; PAC, premature atrial contraction; PAT, paroxysmal atrial tachycardia; PVCs, premature ventricular contractions; EKG, electrocardiogram; NOS, not otherwise specified; ABGS, arterial blood gases; HCT, hydrochlorothiazide; HGB, hemoglobin.  
© Copyright by Susan D. Horn, PhD. All rights reserved. Do not quote, copy, or cite without permission.

it represents a composite view of all derangements over the entire hospitalization rather than a picture of a single day or moment. All patients who die during hospitalization are assigned an overall maximum severity of 4.

Discharge severity involves the last calendar day plus the preceding 24 hours. All diagnoses are considered to determine whether each has resolved or returned to a low severity level. All

deaths are also assigned an overall discharge severity of 4.

According to the CSI's developers, these different scores have different potential applications. Admission severity provides risk adjustment for poor outcomes (for example, death), although in performing such adjustments it is important to remember that all deaths during the admission period are automatically assigned

Table 4. The Computerized Severity Index Matrix for Congestive Heart Failure

Heart Failure							
	428.0-428.9	164.1	212.7	398.91-398.99	506.1	514	518.4
	68.10-668.14	669.40-669.94	996.00-996.09	996.88-996.84	997.1		
Category	1	2	3	4			
ADLS†	• Requires assistance	• Complete dependence					
Cardiovascular	• CO >=2.8L • Edema NOS; JVD <=2cm; CVP <=12cmH2O • Pulse rate <100; ST segment changes on EKG • Palpitations • S4 • Bounding peripheral pulses	• CO 2.7-2.8L • 1+-2+ edema; JVD 3-5cm; CVP 13-15cmH2O • Pulse rate 100-129; PACs, PAT, PVs on EKG • S3	• CO 2.2-1.9L • 3+-4+ edema; JVD 6-9cm; CVP 16-24cmH2O • Pulse rate >=130 • All pulses thready	• CO <=1.8L • Anasarca; JVD >=10cm; CVP >=25cmH2O • Life-threatening arrhythmias or hypotension • All pulses absent			
Genitourinary	• UO >1000cc	• UO 999-500cc	• UO 499-100cc	• UO <=99cc			
Labs ABGS	• pH 7.35-7.45	• pH 7.46-7.50 7.34-7.25	• pH 7.51-7.60 7.24-7.10	• pH >=7.61 <=7.09; pO2 <=50			
Neurology		• Chronic confusion	• Acute confusion	• Unresponsive			
diology rest	• Cardiac enlargement	• Increased cardiothoracic ratio	• Cardiothoracic ratio >=55%; pulmonary edema				
Respiratory		• Dyspnea on exertion; PND; rates <=50%/ <=3 lobes; decreased breath sounds <=50%/ <=3 lobes	• Dyspnea at rest; rates >50%/ >3 lobes; decreased breath sounds >=50%/ >3 lobes	• Apnea; Absent breath sounds >50%/ >3 lobes			
	• White, thin, mucoid sputum	• Hemoptysis NOS	• Cyanosis				

† ADLS, activities of daily living; CO, cardiac output; NOS, not otherwise specified; JVD, jugular venous distention; CVP, central venous pressure; ST, stress test; UO, urinary output; PND, paroxysmal nocturnal dyspnea.  
© Copyright 1986 by Susan D. Horn, PhD. All rights reserved. Do not quote, copy, or cite without permission.

scores of 4. The admission score also is designed to call into question the appropriateness of admission in certain instances—for example, medical admissions with overall admission severity of 1. Maximum severity intends to predict resource need. The 1988 New Jersey experiment—which looked at hospitalization costs—used overall admission severity to explore the adjustment of DRG-based payment levels.<sup>23</sup> Comparisons of admission and

maximum severity could indicate whether severity worsened over the hospital stay—a potential indicator of substandard care. Finally, discharge severity could suggest instances of premature discharge.

The CSI review process is initiated by entering a patient's ICD-9-CM diagnostic codes into a personal computer with an enhanced core memory space. The software then asks a series of questions about the clinical variables

associated with the relevant severity matrices (for example, "What are the highest and lowest temperatures?" and "What are the highest and lowest serum potassium levels?"). Data can be drawn from all notes in the medical record, including countersigned notes of trainees. The order of the questions can be varied depending on the organization of the medical record—for example, all questions concerning vital signs could be asked first. The data entry process includes range checks so that grossly aberrant or impossible values cannot be entered.

Hospital-specific normal ranges are also entered to guide computations involving selected continuous variables. If a computer is not available at the review site, paper forms containing all queries and answer spaces as they appear on the computer screens can be created elsewhere—for example, from discharge abstract computer files. The abstracted data would subsequently be entered into the computer.

#### Clinical Assessment of the CSI

*The use of ICD-9-CM.* The CSI's dependence on ICD-9-CM diagnostic codes is simultaneously an important strength and a potential Achilles heel. On the benefit side, the use of ICD-9-CM means that the CSI can be easily integrated into the many functions within the health care sector that rely on administrative data and ICD-9-CM coding. Up front, the CSI attempts to quiet widespread concerns about the accuracy of ICD-9-CM coding<sup>21-24</sup> by using its criteria to validate the presence of coded conditions. This is a potentially useful endeavor, although it focuses exclusively on overcoding. However, because most of the level 1 criteria in the severity matrices intentionally represent normal to very mildly deranged values of routine laboratory tests, it is not clear that the CSI's current approach will fulfill its objective. For example, severity level 1 criteria for acute duodenal ulcer with obstruction, hemorrhage, and perforation (code 532.21) include continuous variables in largely normal ranges, such as a hematocrit value greater than or equal to 30%, on arterial pH value of 7.35 to 7.45, and a serum potassium value of

3.5 to 5.2 mEq/L. Thus, code 532.21 could be validated in patients with a normal hematocrit or serum potassium value, even if acute duodenal ulcer is not present.

Another practical concern relating to ICD-9-CM involves the use of discharge diagnoses for admission reviews. The failure to specify which ICD-9-CM codes pertain to the admission period could potentially result in assignment of disease-specific severity scores to conditions not present during this period—for example, if a patient is admitted to the hospital for diverticulitis of the colon (code 562.11) and three days later develops septicemia caused by *Escherichia coli* (code 38.42). The severity matrices for these two conditions contain multiple identical criteria, such as pulse, digestive symptoms (nausea and vomiting), fever, high white blood cell count, and hematocrit. If multiple, severe criteria are identified, both codes could receive severity scores greater than one and could thereby contribute to the overall admission severity, although the septicemia was not present upon admission. This problem can be solved by flagging those ICD-9-CM codes present upon admission, but this necessitates an extra step during the review process and is not readily applicable to situations in which administrative data are used to produce the CSI abstraction forms. In addition, "diagnoses" listed as present on admission are often revised as diagnostic information becomes available or as the clinical course evolves. This concern also argues against use of the CSI concurrently during hospitalization.

The CSI appropriately attempts to deal with the redundancy of much of the ICD-9-CM lexicon by prohibiting multiple codes from the same organ system to be used in producing overall severity scores. To facilitate this process, the CSI groups many of the ICD-9-CM codes into 14 related disease groups. In computing overall severity, the CSI selects the single diagnosis within a related disease group that has the highest severity and ignores all other related diagnoses. Some of the related disease groups are very expansive and encompass

many pathophysiologic processes if they occur in the same organ system. For example, lung cancer and pneumonia are in the same related disease group, as are AMI and malignant hypertension. This represents a fairly rigorous protection against the excessive ICD-9-CM coding or "creep" that may have occurred as a result of the implementation of DRG-based payment.<sup>25-27</sup>

The first related disease group raises the most questions because it contains approximately 1,000 codes representing a variety of conditions—most eye diseases are included, as well as selected malignancies, diabetes mellitus, and renal problems. For example, viral conjunctivitis, malignancy of the male genital tract, bladder, and kidney, all manifestations of diabetes mellitus, papilledema, strabismus, acute and chronic renal failure, and surgical complications of the urinary tract are all included. Because of the breadth of this disease group and other related disease groups, the question arises as to whether this relatively vigorous strategy against the possibility of code creep is *too* stringent.

Although much of the CSI computer algorithm is devoted to minimizing the effect of excessive coding on overall severity scores, the system nonetheless requires complete ICD-9-CM coding. The clinical criteria on the severity matrices do not generally include indications of complications outside the organ system in which the disease originates. For example, if a patient develops a stroke or acute renal failure as a result of an AMI, each of these conditions must receive its own ICD-9-CM code—the AMI severity matrix does not contain information concerning these complications. Some examples are more subtle—for instance, low body temperature is not listed on the pneumonia severity matrix, although it has been shown to be an important predictor of risk of death for this condition.<sup>28</sup> The CSI relies on the ICD-9-CM coding to capture this finding.

This requirement of "complete" ICD-9-CM coding is a potential problem because of the usual restriction in administrative data bases of diagno-

sis coding fields to five. For example, there is some compelling evidence that Medicare's truncation of diagnosis spaces at five biases against the reporting of some chronic illnesses among patients who die.<sup>29</sup> Coding of such clinical signs as hypothermia is probably likely to be even more variable. Medicare plans to expand the number of diagnosis fields on its billing form (UB-82) to nine as of April 1, 1992,<sup>30</sup> but it is not yet clear whether this action will actually increase the completeness of the diagnosis coding upon which the CSI depends. In California, where coding is permitted up to 25 diagnoses, certain conditions are still underreported.<sup>31</sup>

*Content of CSI severity matrices.* The clinical logic of the CSI is transparent, which renders the CSI easily accessible from the severity matrices associated with the approximately 820 disease groups. Since the matrices are disease specific they offer the ready opportunity for physicians using the CSI to debate the matrices' merits in explicit clinical terms. Most criteria requested on the CSI matrices reflect acute manifestations of diseases, although a few chronic findings and a gross assessment of activities of daily living are sometimes included (Table 4).

The CSI's developers attempted to rely on objective data elements, eschewing treatment variables, to classify severity. Consequently the CSI focuses on physical examination findings, signs and symptoms, laboratory results, and routine radiologic and other test results. Approximately one-half of the CSI's factors represent vital signs or laboratory findings that are generally reliably abstracted from the medical record in the context of other severity measurement systems, such as APACHE and MedisGroups.<sup>32</sup> However the remainder tend to be more descriptive and qualitative, and they can be grouped into the following three categories: (1) items that physicians often measure unreliably; (2) items that depend on patient report; and (3) descriptive clinical characteristics that might be abstracted unreliably from the medical record.

A number of factors represent items

## A Description and Clinical Assessment of the Computerized Severity Index

that are notorious for interobserver disagreements and high variability across physicians, including Kussmaul respiration, pulsus paradoxus, pericardial friction rub, delayed reflex relaxation, jugular venous distention, and specified chest radiographic findings (for example, moderate pulmonary vascular congestion).<sup>43-46</sup>

Does this compromise the utility of the CSI? It is important to note that other severity measures, such as MedisGroups, use certain findings that are similarly subject to interphysician disagreements.<sup>4</sup> Because of their potential unreliability, MedisGroups assigns low weight to many of these findings and attempts to improve objectivity by often requiring evidence of findings on specified tests that can be expensive and technologically sophisticated (for example, computerized tomography and endoscopy). The CSI specifically attempts to avoid dependence on expensive testing, while relying heavily on the more descriptive documentation from the physician or nurse.

Many of these clinically unreliable findings are nonetheless implemented by physicians as indicators of severity during actual patient care. This relationship to the patient care context may enhance the appeal of the CSI to physicians. However the lingering concern is the potential for "gaming" (manipulating the system). This is a possibility, especially given the weight accorded to qualitative, descriptive findings. In addition, it is impossible to evaluate retrospectively the accuracy of the clinical examination findings. Nonetheless, the need to have two criteria at a given severity level, the complexity of the system, and other features of the computer algo-

rithm may diminish the likelihood of significant reward from gaming attempts.

Numerous items require patient reporting and thus may be subject to the vagaries of recall and differences in pain or discomfort thresholds regarding generalized weakness, fatigue, constipation, moderate and severe headache, dry mouth, numbness and tingling, and uncomfortable feeling in the throat during respiration. Many of the differences in patient reporting these types of factors would probably vary randomly across providers. However certain systematic biases could possibly derive from socioeconomic, cultural, and language characteristics of a patient population. Many of these criteria are assigned to severity level 1 ("normal to mild") and therefore will not greatly influence severity scores, but some do have a larger impact.

Finally, a number of factors are based on colloquial and informal language—for example, "large, flabby protuberant abdomen" in cystic fibrosis; "waddling gait" in Vitamin D deficiency; and "stuffy nose" in influenza with pneumonia. Many such factors are assigned to severity level 1, but some are given higher assignments. This is not a common problem but it is of concern because most such vernacular terms are not addressed in the CSI Help File. Abstraction guidelines specify that the description must be written verbatim in the medical record; however the required semantics are often not standard medical terminology.

*Process of reviews.* As with other severity measures such as MedisGroups, the CSI derives from the worst derangements during the designated review period. This raises concern about the potential impact of iatrogenic events and quality shortfalls on the severity score. In the case of maximum severity, the possible role of substandard quality in influencing scores is recognized by the CSI's developers and is part of the basis for their quality assessment strategy (for example, comparing admission and maximum severity). However the admission score may also be influenced by quality problems: the admission period (from 25 to 48 hours) may be a crucial time for certain conditions.

The time frame for data collection concerning surgery and/or other procedures is fairly narrow. No information is permitted from operating room documentation, and all signs and symptoms from the first postoperative hour are ignored. Certain specified complications in the immediate postoperative period are not considered, but only for a short time. For example, confusion and disorientation are not recorded up to only 12 hours after brain surgery, and cardiac output and arterial blood gas findings are not considered up to only 24 hours after open heart surgery. It is possible that these narrow windows will allow normal and time-limited postoperative derangements to be gathered and treated as severe.

### Discussion

The CSI is a methodology with many facets, including severity matrices for over 820 conditions, weighting rules which consider the differential impact of the gamut of principal and secondary diagnoses, and various strategies for data collection to respond to multiple policy-oriented goals. As such, gaining an in-depth appreciation for the detail of the CSI is a large task, and this article merely represents a first step.

Little research has yet been published concerning the performance of the CSI, although two recent publications provide some insight into its performance for predicting resource use.<sup>22,23</sup> Using data from the New Jersey experiment, McGuire<sup>22</sup> found that the CSI performed better than several models derived from various permutations of the DRG classification system in terms of reducing variance in hospitalization costs. However McGuire did not detail the cost of CSI data collection, and as of this time, New Jersey has no active plans to employ the CSI in setting hospital reimbursement levels. Horn<sup>23</sup> and colleagues used data from 2,378 of the 15,000 cases from the five university hospitals demonstration, finding that DRGs adjusted for maximum CSI scores explained 54% of variation in length of stay. They also found that admission CSI score was strongly predictive of in-hospital mortality, although it is important to recall that patients

who die within the admission period are all assigned scores of 4. One important study<sup>17</sup> published by independent investigators concluded that the maximum CSI score was a better predictor of hospitalization costs than admission severity measures or measures based on discharge abstract data.

Even less has been published concerning the relationship between CSI scores and quality of care. A forthcoming publication<sup>18</sup> examines the association between CSI trajectory information (a comparison between admission and maximum scores to determine if the patients' conditions worsened or remained unchanged) and quality of hospital care for patients with AMI or patients undergoing coronary artery bypass graft surgery. For AMI patients, a worsening trajectory was associated with higher rates of potential quality problems. However for coronary artery bypass graft patients, the association varied depending on how quality was assessed. Clearly, this is an area that requires further investigation.

Thus it is important for additional research to focus on the application of the CSI—its ability to fulfill its stated goals, such as the prediction of resource use or its utility as a screen for the quality of hospital care.

#### Areas for Further Study

First, the dependence upon ICD-9-CM prompts a number of questions. Of particular importance is the practical issue of the source of the ICD-9-CM codes to be used in the various reviews. If all codes from the discharge abstract are used, admission reviews may result in the paradoxical assignment of a severity score to a condition that does not exist upon admission. In general, enough provisions against double counting exist to prevent such scores from unduly influencing the overall severity rating. However sometimes these protections may be insufficient and a diagnosis that is not present upon admission could affect the overall admission score. (Whether this is an important concern remains to be studied.) In addition, the approach toward validating the ICD-9-CM codes must be examined. Given its reliance on nor-

mal values, the current strategy could result in "validation" of potentially inaccurate codes; however the extent of this problem is as yet unknown.

Second, the clinical content of the severity matrices must be further evaluated by physicians, especially those affected by the CSI in their hospital or state. The forming clarity of the severity matrices will facilitate this review, and much of the final decisions about the validity of the criteria may be a matter of individual clinical judgment. In this review, it is important to note that a relatively narrow definition of "disease" was employed, such that many complications outside the organ system in which the disease originated (for example, distant metastases in malignancy) are not generally included on the matrices. Whether standard ICD-9-CM coding practices are sufficiently complete to fill this void remains to be seen.

Third, many of the CSI's clinical criteria are descriptive and qualitative. Some represent items that physicians measure with tremendous variability, others involve characteristics requiring patient report and judgment, while still others are described by idiomatic or vernacular terms and may be abstracted unreliably from the medical record. Should the CSI be used for policy-related purpose, the most pressing issue is whether the system can be gamed. An additional question is whether documentation biases across hospitals or physicians could artificially affect scores.

Fourth, the sensitivity of CSI scores to quality shortfalls should be explored. Chances are that the maximum score is affected by substandard care, but whether this makes it an inappropriate tool for determination reimbursement is open to debate—the DRGs set a compelling precedent of a patient classification system used for payment purposes in which the resulting patient classifications can sometimes reflect quality problems. The more important question involves the admission score and its potential use as a risk adjuster for quality assessment studies. The use of the worst values from the first two hospital days could possibly include some findings related to iatrogenic

events and quality problems.

Fifth, the process of CSI reviews and the time required should be examined—particularly the maximum severity, which depends on the worst values from the entire hospitalization period and may therefore necessitate a lengthy review for certain types of cases. Although, according to the CSI's developers, an average of 32 questions are asked per case, the entire medical record must be scanned. In addition, for each case a somewhat different set of questions is asked depending on the constellation of diagnoses. Thereby whether this significantly affected reviewer training or the review process slowed, especially in tertiary teaching hospitals (which often have voluminous documented medical records), remains to be studied. The costs of widespread data collection and its burden on medical record departments must be evaluated.

Finally, the predictive validity of the CSI must be explored—does the system do what it intends to do? For example, does it predict resource use, and is it a good detector of substandard care or premature discharge? Preliminary findings concerning cost predictions have come from the New Jersey experiment, but more widespread application is necessary for generalizable results. Once this research has been performed it will be possible to make a more informed and thoughtful judgment concerning the CSI.

#### References

1. Iezzoni LI, Shwartz M, Restuccia J: The role of severity information in current health policy debates: A survey of state and regional concerns. *Inquiry* 28:117-128, 1991.
2. Jencks SJ, Dobson A: Refining case-mix adjustment: The research evidence. *N Engl J Med* 317:679-686, 1987.
3. 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* 3:482-490, 1988.
4. Iezzoni LI: Measuring the severity of illness and case mix. In Goldfield N, Nash DB (eds): *Providing Quality Care: The Challenge to Clinicians*. Philadelphia: American College of Physicians, 1989, pp 70-105.

# A Description and Clinical Assessment of the Computerized Severity Index

5. Kaple JG: Using severity indices to assess quality of care. *Business and Health* 5:23-28, 1987.
6. Aquilina D, McLaughlin B, Levy S: Using severity data to measure quality. *Business and Health* 5:40-42, 1988.
7. Iezzoni LI: Severity standardization and hospital quality assessment. In Couch JB (ed): *Health Care Quality Management for the 21st Century*. Tampa, FL: The American College of Physician Executives, Hillsboro Printing Company, 1991, pp 177-234.
8. Knaus WA, et al: APACHE II: A severity of disease classification system. *Crit Care Med* 13:818-829, 1985.
9. Wagner DP, Knaus WA, Draper EA: Physiologic abnormalities and outcome from acute disease: Evidence for a predictive relationship. *Arch Intern Med* 146:139-1396, 1986.
10. Knaus WA, et al: An evaluation of outcome from intensive care in major medical centers. *Ann Intern Med* 104:410-418, 1986.
11. Gonnella JS, Hornbrook MC, Louis DZ: Staging of disease: A case-mix measurement. *JAMA* 251:637-644, 1984.
12. Conklin JE, et al: Disease staging: Implications for hospital reimbursement and management. *Health Care Financing Review* (annual supplement) 13-21, 1984.
13. Brewster AC, et al: MEDISGRPS: A clinically based approach to classifying hospital patients at admission. *Inquiry* 12:377-387, 1985.
14. Iezzoni LI, Moskowitz MA: A clinical assessment of MedisGroups. *JAMA* 260:3159-3163, 1988.
15. Iezzoni LI, et al: Admission and mid-stay MedisGroups scores as predictors of death within 30 days of hospital admission. *Am J Public Health* 81:74-78, 1991.
16. Young WW: Incorporating severity of illness and comorbidity in case-mix measurement. *Health Care Financing Review* (annual supplement) 23-31, 1984.
17. Young WW, Swinkola RB, Zorn DM: The measurement of hospital case mix. *Med Care* 20:501-512, 1982.
18. Horn SD, Horn RA: The computerized severity index: A new tool for case-mix management. *J Med Systems* 10:73-78, 1986.
19. Horn SD: Physician profiling: How it can be misleading and what to do. *Consultant* 27:86-94, 1987.
20. Horn SD, Backofen JE: Ethical issues in the use of a prospective payment system: The issue of a severity of illness adjustment. *J Med Philo* 12:145-153, 1987.
21. Iezzoni LI, Moskowitz MA, Daley J: *A Description and Clinical Assessment of the Computerized Severity Index*. Health Policy Research Consortium (unpublished report). Prepared for the Health Care Financing Administration under Cooperative Agreement No. 18-C-98526/1-05. Brandeis University, Waltham, MA, July, 1989.
22. Horn SD, et al: The relationship between severity of illness and hospital length of stay and mortality. *Med Care* 29:305-317, 1991.
23. McGuire TE: An evaluation of diagnosis-related group severity and complexity refinement. *Health Care Financing Review* 12:49-60, 1991.
24. Horn SD: Validity, reliability and implications of an index of inpatient severity of illness. *Med Care* 19:354-362, 1981.
25. Horn SD, et al: Severity of illness within DRGs: Homogeneity study. *Med Care* 24:225-235, 1986.
26. Horn SD, et al: Interhospital differences in severity of illness: Problems for prospective payment based on diagnosis-related groups (DRGs). *N Engl J Med* 313:20-24, 1985.
27. Horn SD, et al: Severity of illness within DRGs: Impact on prospective payment. *Am J Public Health* 75:1195-1199, 1985.
28. Schumacher DN, et al: Severity of illness index and the adverse patient occurrence index: A reliability study and policy implications. *Med Care* 25:695-704, 1987.
29. Snee VN: The International Classification of Diseases: 9th Revision (ICD-9). *Ann Intern Med* 88:424-426, 1978.
30. Iezzoni LI, Moskowitz MA: Clinical overlap among medical diagnosis-related groups. *JAMA* 255:927-929, 1986.
31. Hsia DS, et al: Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *N Engl J Med* 318:352-355, 1988.
32. McMahon LF, Smits HL: Can Medicare prospective payment survive the ICD-9-CM disease classification system? *Ann Intern Med* 104:562-566, 1985.
33. Lloyd SS, Rissing JP: Physician and coding errors in patient records. *JAMA* 254:1330-1336, 1985.
34. Iezzoni LI, et al: Coding of acute myocardial infarction: Clinical and policy implications. *Ann Intern Med* 109:745-751, 1988.
35. Simborg DW: DRG creep: A new hospital-acquired disease. *N Engl J Med* 304:1602-1604, 1981.
36. Health Care Financing Administration, U.S. Department of Health and Human Services: Medicare program: Changes to the inpatient hospital prospective payment system and fiscal year 1986 rates: Final rule. *Federal Register*. 50:35700, Sep 3, 1985.
37. Steinwald B, Dummit LA: Hospital case-mix change: Sicker patients or DRG creep? *Health Aff* 8:35-47, 1989.
38. Daley J, et al: Predicting hospital-associated mortality for Medicare patients with stroke, pneumonia, acute myocardial infarction, and congestive heart failure. *JAMA* 260:3617-3624, 1988.
39. Jencks SF, Williams DK, Kay TL: Assessing hospital-associated deaths from discharge data: The role of length of stay and comorbidities. *JAMA* 260:2240-2246, 1988.
40. Federal Register: Rules and Regulations. 58(169):43213-43215, Aug 30, 1991.
41. Meux EF, Stith SA, Zach A: Report of results from the OSHPD reabstracting project: An evaluation of the reliability of selected patient discharge data July through December 1988 (unpublished report). Sacramento, CA: Patient Discharge Data Section, Office of Statewide Health Planning and Development, State of California, Dec 1990.
42. Thomas JW, Ashcraft MLF: Measuring severity of illness: A comparison of interrater reliability among severity methodologies. *Inquiry* 26:483-492, 1989.
43. Eddy DM: Variations in physician practice: The role of uncertainty. *Health Aff* 3:74-89, 1984.
44. Wiener S, Nathanson M: Physical examination: Frequently observed errors. *JAMA* 236:852-855, 1976.
45. Gjørup T, Bugge PM, Jensen AM: Interobserver variation in assessment of respiratory signs. *Acta Med Scand* 216:61-66, 1984.
46. Johnson JE, Carpenter JL: Medical house staff performance in physical examination. *Arch Intern Med* 146:937-941, 1986.
47. Thomas JW, Ashcraft MLF: Measuring severity of illness: Six severity systems and their ability to explain cost variations. *Inquiry* 28:39-55, 1991.
48. Iezzoni LI, et al: The utility of severity of illness information in assessing the quality of hospital care: The role of the clinical trajectory. *Med Care*, in press, 1992.





Faxed 2/16/99

THE HOSPITAL & HEALTHSYSTEM ASSOCIATION OF PENNSYLVANIA

RECEIVED

Carolyn F. Scanlan  
President and Chief Executive Officer

99 FEB 18 AM 8:48

INDEPENDENT REGULATORY  
REVIEW COMMISSION

February 16, 1999

Mr. John R. McGinley, Jr.  
Chairman  
Independent Regulatory Review Commission  
Commonwealth of Pennsylvania  
333 Market Street, 14th Floor  
Harrisburg, Pennsylvania 17101

ORIGINAL: 1995  
MIZNER  
Fax received 2/17/99  
COPIES: de Bien  
Harris  
Sandusky  
Legal

Re: 28 PA. Code Chapters 911 and 912

Dear Mr. McGinley:

The Hospital & Healthsystem Association of Pennsylvania (HAP), on behalf of its members (more than 225 acute and specialty hospitals and health systems in the commonwealth), appreciates the opportunity to comment on the Pennsylvania Health Care Cost Containment Council's proposed rulemaking (published in the *Pennsylvania Bulletin* on January 16, 1999) amending the council regulations.

The current regulation specifies a particular methodology to evaluate the effectiveness of patient care. That methodology was selected based on available systems in 1987. By specifying a particular methodology, the council is precluded from selecting a different vendor and/or methodology that may be more effective and economical. As HAP understands the proposed amendments, their purpose is to give the council the flexibility to utilize a different vendor if it appears that a more effective and economical system is available. It also gives the council the opportunity to rapidly seek another vendor and/or methodology if the current vendor (MediQual) fails to perform. Based upon this understanding of the intent of the proposed amendments, HAP supports the proposed rulemaking.

Two relevant issues regarding the proposed rulemaking require specific attention by the council, the Independent Regulatory Review Commission, and the legislature. The council needs flexibility to select a patient severity methodology which allows the council to measure the effectiveness of health care providers. Additionally, the potential impact of the proposed rulemaking needs to be better understood. The following observations/ recommendations address these two issues.

4750 Lindle Road  
P.O. Box 8600  
Harrisburg, PA 17105-8600  
717.564.9200 Phone  
717.561.5334 Fax  
cscanlan@hap2000.org



Mr. John McGinley, Jr.  
February 16, 1999  
Page 2

### **Patient Severity**

To ensure that the council has the needed flexibility to select a more effective and economical severity methodology, it is important that the amendments permit the council to consider the full array of severity adjustment systems currently available and which may be developed in the future. The proposed change in the definition of "patient severity" could prove to limit the council's flexibility in selecting an alternative methodology, even if the alternative methodology provides better and more complete information on the effectiveness of health care providers.

**HAP recommends that the council modify the definition of "patient severity" to afford greater flexibility in selecting severity methodologies. HAP recommends the following definition of "patient severity" be used by the council in final rulemaking:**

*Patient severity*—A measure of severity of illness as defined by the council through the application of either: 1) a reputable discharge abstract-based severity system using appropriate indicators (i.e., diagnosis, treatments, demographics, and resource utilization) from the standard patient discharge abstract; or 2) a reputable severity system using data (i.e., diagnosis, treatments, demographics, and other relevant factors) abstracted from individual patient records.

### **Potential Economic or Fiscal Impact**

Adoption of the proposed rulemaking, in and of itself, will have no fiscal impact. The proposed amendments will not, per se, impose additional paperwork requirements. However, when the council chooses to exercise the flexibility afforded it through the proposed rulemaking there is the potential for significant impact (positive and/or negative).

Currently, hospitals incur significant costs and paperwork requirements associated with collecting data using the mandated MediQual system (estimated at \$40 million to \$50 million annual cost for all Pennsylvania hospitals). These costs include the fees paid by hospitals to MediQual to license the mandated severity adjustment system as well as the cost for personnel to manually complete MediQual patient abstract forms for approximately 1.3 million inpatient discharges per year, enter the abstracted data into the proprietary MediQual software, transmit the abstracted data to MediQual, and validate

Mr. John McGinley, Jr.

February 16, 1999

Page 3

the abstracted data. Additionally, the difficulties encountered in implementing MediQual's Atlas 2.0 software illustrate the potential for unnecessary burdens on hospitals. The costs of the MediQual mandate are in addition to the costs hospitals incur in creating standard patient discharge abstract data sets that are submitted directly to the council for all inpatient discharges and all ambulatory surgery cases.

Two of the council's stated ongoing objectives are "to make data collection more effective and to potentially reduce costs incurred by reporting providers." Consequently, if the council elects to adopt a different methodology and/or vendor, the cost of compliance should be a principal consideration. The council should also consider the benefits to the commonwealth of adopting a different methodology. If an alternative methodology provides better information on the effectiveness of health care providers, all Pennsylvania residents, their insurance companies and/or their employers could make better choices on selecting providers. Improved information on the health care market and potentially lower costs of compliance could spawn a more competitive market which could improve the quality of care at a lower cost.

**HAP recommends that following adoption of the final-form publication of revised amendments (see patient severity above) the council exercise due diligence in exploring alternative severity adjustment systems for possible adoption. Due to the potential direct impact on the regulated community, the council should seek involvement of the regulated community in the selection of a severity adjustment methodology.**

In summary, HAP supports the intent of the proposed regulations and believes that they can be improved by enhancing the definition of "patient severity" to allow the council greater flexibility in evaluating severity adjustment methodologies. Additionally, HAP recommends that, upon adoption of the proposed rulemaking, the council exercise due diligence in evaluating severity adjustment systems with the active participation of the regulated community.



Mr. John McGinley, Jr.  
February 16, 1999  
Page 4

HAP is committed to improving the timeliness, quality, and effectiveness of data reported to and by the council. We believe that the proposed rulemaking is an important step in reaching this objective. We offer our cooperation and assistance in whatever capacity is needed. If you have any questions, or if we can be of further assistance, please feel free to call me at (717) 561-5314 or Martin Ciccocioppo at (717) 561-5363.

Sincerely,

A handwritten signature in cursive script that reads 'Carolyn F. Scanlan'.

CAROLYN F. SCANLAN  
President and Chief Executive Officer

CFS/mjc

January 1999

# Accuracy of Risk-Adjusted Mortality Rate As a Measure of Hospital Quality of Care

ORIGINAL: 1995

WIZNER

COPIES: de Bien  
Harris  
Sandusky,

RECEIVED

88 of 23 AM 11:54

INDEPENDENT LABORATORY  
REVIEW COMMISSION

J. WILLIAM THOMAS, PHD,\*† AND TIMOTHY P. HOFER, MD, MST‡

**OBJECTIVES.** Reports on hospital quality performance are being produced with increasing frequency by state agencies, commercial data vendors, and health care purchasers. Risk-adjusted mortality rate is the most commonly used measure of quality in these reports. The purpose of this study was to determine whether risk-adjusted mortality rates are valid indicators of hospital quality performance.

**METHODS.** Based on an analytical model of random measurement error, sensitivity and predictive error of mortality rate indicators of hospital performance were estimated.

**RESULTS.** The following six parameters were shown to determine accuracy: (1) mortality risks of patients who receive good quality care and (2) of those who receive poor quality care, (3) proportion of patients (across all hospitals) who receive poor quality care, (4) proportion of hospitals considered to be "poor quality," (5) patients' relative risk of receiving poor quality care in "good quality"

and in "poor quality" hospitals, and (6) number of patients treated per hospital. Using best available values for model parameters, analyses demonstrated that in nearly all situations, even with perfect risk adjustment, identifying poor quality hospitals on the basis of mortality rate performance is highly inaccurate. Of hospitals that delivered poor quality care, fewer than 12% were identified as high mortality rate outliers, and more than 60% of outliers were actually good quality hospitals.

**CONCLUSIONS.** Under virtually all realistic assumptions for model parameter values, sensitivity was less than 20% and predictive error was greater than 50%. Reports that measure quality using risk-adjusted mortality rates misinform the public about hospital performance.

**Key words:** outcome and process assessment; quality of health care; hospitals; mortality; hospital mortality; health services research. (Med Care 1999;37:83-92)

The science of mortality risk-adjustment has improved dramatically since the Health Care Financing Administration (HCFA) released its first hospital mortality data report in 1986. More than a dozen methodologies are now available to estimate with reasonable accuracy probability of death as a function of individual patient demographic (eg, age, sex) and clinical characteristics (eg, diagnoses, laboratory findings, vital signs). With good risk-adjustment, there are only two

factors to which differences between a hospital's observed mortality rate and its expected rate can be attributed: random variation and quality of care. When a hospital's observed mortality rate is so much greater than its expected rate that the difference is considered unlikely to have occurred by chance, the hospital is termed a high outlier and is presumed to be delivering poor quality care.

Is this presumption correct? In one of the few studies that have attempted to address this issue,

\*From the Department of Health Management and Policy, School of Public Health, University of Michigan, Ann Arbor, Michigan.

†From the Department of Veterans Affairs, Center for Practice Management & Outcomes, HSR&D, Ann Arbor, Michigan.

‡From the Department of Internal Medicine, School of Medicine, University of Michigan, Ann Arbor, Michigan.

Address correspondence to: J. William Thomas, PhD, Department of Health Management and Policy, School of Public Health, The University of Michigan, Ann Arbor, MI 48109; e-mail: jwthomas@umich.edu.

Received January 26, 1998; initial review completed February 26, 1998; accepted June 10, 1998.

Park et al<sup>1</sup> examined medical records for samples of acute myocardial infarction (AMI) and congestive heart failure (CHF) patients treated in hospitals that had been identified as high mortality rate outliers and for samples of similar patients treated in nonoutlier facilities. Each chart was assigned a quality-of-care score based on the documented process of care and a severity of illness score based on vital signs and physiologic findings. With simulation analyses, it was determined that, depending on condition, 56% to 82% of the observed mortality rate differences between high-outlier and nonoutlier facilities were attributable purely to random binomial variation. The principal finding of the study, however, was that hospitals identified "with unexpectedly high age-sex-race-disease-specific death rates *do not* (emphasis added) provide lower quality of care than do untargeted hospitals."<sup>1</sup>

This negative finding—that quality of care in high outlier hospitals is not worse than in nonoutlier hospitals—has done little to discourage use of mortality rate as an indicator of hospital quality of care. Although HCFA suspended its mortality data releases in 1993, other organizations—state data agencies, business coalitions, commercial health data vendors, the media—now produce an increasing number of comparative performance reports for hospitals. Most of the reports utilize risk-adjusted mortality rate as the principal measure of quality of care.<sup>2</sup> Like the HCFA's first data release, these contemporary performance reports continue to be controversial, with critics charging that the figures are inaccurate and misleading.<sup>3,4</sup>

The most common criticism still centers on the adequacy of risk-adjustment methodologies. Although the adequacy of risk-adjustment is still being debated, in this article we will ignore the risk-adjustment issue and instead focus on the following questions:

If all influences of casemix and severity differences among providers were to be removed through perfect risk-adjustment, would hospital mortality rates then be able to identify poor quality providers accurately?

It is not possible to answer this question empirically because, in spite of recent advancements, mortality risk-adjustment methodologies remain less than perfect.<sup>2</sup> Further, answering the question empirically would require an independent, valid measure of quality with which poor quality providers could be identified. No currently avail-

able outcome-based measure meets the validity criterion.

In their investigation of the usefulness of mortality rates for identifying poor quality providers, Hofer and Hayward<sup>5</sup> attempted to circumvent the two problems above by simulating mortality experience for a hypothetical set of hospitals, with perfect risk-adjustment and with prior perfect knowledge of the identity of hospitals providing poor quality care. Under the simulated conditions, they found mortality rates to perform poorly for identifying low quality providers. Zalkind,<sup>6</sup> using a Monte Carlo simulation methodology similar to that of Hofer and Hayward,<sup>5</sup> arrived at the same conclusion.

Each of these simulation studies found that the ability to identify poor quality hospitals accurately on the basis of mortality rates varied depending on numbers of treated cases per facility and on differences in quality-associated mortality rates between good and poor quality facilities. In the article presented here, we used an analytic model to derive precise measurements for the accuracy of mortality rate identification of poor quality hospitals. Although our analytic methodology was different from that of Hofer and Hayward<sup>5</sup> and Zalkind,<sup>6</sup> several key assumptions were similar. Like Hofer and Hayward<sup>5</sup> and Zalkind,<sup>6</sup> we assumed no casemix or severity differences across hospitals, and we focused on a hypothetical system of hospitals, a small portion of which deliver very poor quality care. Like Zalkind,<sup>6</sup> we also assumed that all facilities treat exactly the same number of patients. With these assumptions, mortality rate differences across hospitals would be attributable only to quality-of-care differences and to random binomial variation.

## Methods

### Defining Accuracy for Indicators of Provider Quality Performance

What does accurate mean in this context? In answering the question, it must be recognized that the mortality rate observed for a hospital, even when perfectly risk-adjusted, is not likely to be the true rate for that hospital. Certain hospital deaths represent sentinel events—deaths from causes that, in our health care system, should never occur—and even one such death signals an important quality problem<sup>7</sup>; however, few hospital deaths are of this type.

When a provider serves a large population of clinically similar patients, say several thousand, the observed mortality rate is likely to be equal or very close to the true rate for that population. For any specific procedure or condition, however, hospital caseloads are generally not this large; a single hospital might have 400 cases or fewer. To interpret the variability in hospital mortality rates properly, it is necessary to view the patients treated at a hospital as representing a sample drawn randomly from a large population of similar cases. If the mortality probability across the entire patient population is  $P$ , a single sample from the population—especially one that includes few patients—can have an observed mortality rate that is much lower or much higher than  $P$ . Elementary statistical principles tell us that if simple random samples of size  $N$  are drawn repeatedly from the population and if a mortality rate calculated for each, the mean of these rates would indeed equal  $P$ , but the rates observed in individual samples would vary around  $P$  and would be distributed approximately normally\* with a standard deviation of

$$\sigma_P = (P [1 - P])^{1/2}.$$

What do we mean by poor quality? Although often described in terms of "good" or "poor," quality actually encompasses numerous dimensions, and each dimension varies as a continuum rather than as a characteristic that is either present or absent.<sup>8,9</sup> For many of these dimensions, the nature of the relation to patient outcome is not at all clear, and reliable measurements are difficult to obtain. Because of such problems, measurement frequently is limited to construction of some form of simple index, and in relating the index to outcomes, quality usually is dichotomized into ranges like "good" and "poor." For our analyses, we shall consider poor quality hospitals to be those in which patients have relatively high risks of receiving poor quality care, and therefore of suffering poor outcomes; good quality hospi-

tals are those in which patients' risks of receiving poor care is small.

With these concepts, we utilize the following parameters to define accuracy for mortality rate indicators of hospital quality:

- The probability of death for each person in a population of patients, all of whom receive good quality (GQ) care [ $P_{D|GQ}$ ];
- The mortality probability for each person in a population of patients, all of whom receive poor quality (PQ) care [ $P_{D|PQ}$ ];
- The fraction of the total patient population that receives poor quality care [ $P_{PQ}$ ];
- The fraction of hospitals that are providing unusually high levels of poor quality care (PQ hospitals) [ $P_{PQH}$ ];
- The "hospital poor quality ratio"—ratio of the probability of receiving poor quality care for patients treated in poor quality hospitals divided by the probability of receiving poor quality care for patients treated in good quality hospitals [ $HPQR$ ]; this is a measure of relative quality performance for hospitals labeled as poor quality compared with other hospitals; and
- The number of patients treated at each hospital [ $N$ ].

Using these six parameters, we can derive the following (see appendix):  $P_D$ , the probability of death averaged over all patients at all hospitals;

$$P_{D|PQH}.$$

average mortality rate at poor quality hospitals;

$$P_{D|GQH}.$$

average mortality rate for good quality hospitals; and TrimPoint—hospitals having rates above the trim point are labeled outliers and are presumed to be delivering poor quality care.

The issue of accuracy for identification of poor quality hospitals is described graphically in Figure 1. Curve A, the distribution of risk-adjusted mortality rates across all hospitals, is approximately normally distributed with mean  $PD$  and standard deviation

$$(PD[1 - PD]/N)^{1/2}.$$

It represents 100% of hospitals and is the sum of curves B and C, the separate mortality rate distributions for good quality and poor quality hospitals, respectively. These distributions also are

\*The sampling distribution of the binomial is considered sufficiently symmetric to be approximated by the normal distribution if both  $N P$  and  $N (1 - P)$  are greater than 5, or equivalently if  $P \pm \sigma_P$  are in the range 0 to 1. [See Loether HJ, McTavish DG. Descriptive and inferential statistics. Boston, MA: Allyn and Bacon, Inc., 1974:409 and Mendenhall W, Beaver RJ. Introduction to probability and statistics. Boston, MA: PWS-Kent Publishing, 1991:217.]

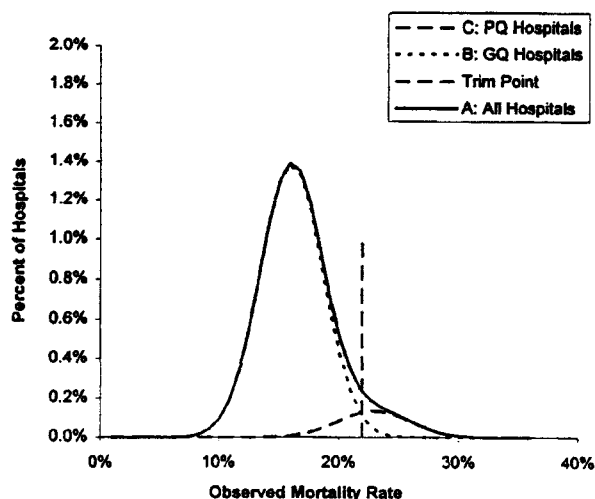


FIG. 1. True positives, true negatives, false positives, and false negatives for mortality rate as an indicator of hospital quality.

approximately normal, with means  $P_{D|GQH}$  and  $P_{D|PQH}$ . Note that since we have no means of differentiating good and poor quality hospitals, curves B and C are actually unobservable. The usual convention is for TrimPoint to be set at the 95th or 97.5th percentile of the sampling distribution, corresponding to a one-tailed or two-tailed test, respectively, of the hypothesis that, for any specific hospital, the hospital's risk-adjusted mortality rate is not statistically greater than the average rate  $P_D$ . With this approach, the set of hospitals represented by the area under curve A and to the right of line TrimPoint are, by definition, high outliers and are considered possibly poor quality providers. The set of hospitals to the right of TrimPoint and under curve C are true positives, and hospitals to the right of TrimPoint and under curve B are false positives. For hospitals to the left of TrimPoint, all of which are considered to be good quality, we have the same issue. Those represented by the area under curve B are true negatives, and those under curve C are false negatives. Thus, the accuracy of a risk-adjusted mortality rate indicator of quality is given by its:

- Sensitivity: the proportion of poor quality hospitals that are high outliers, ie, correctly identified as being poor quality; and
- Predictive error: the proportion of high outliers that are good quality hospitals, ie, incorrectly labeled as delivering poor quality care.

## Estimating Accuracy of Mortality Rate Indicators of Hospital Quality Performance

In this article, our purpose was to determine the potential accuracy of risk-adjusted mortality rate as an indicator of hospital quality in ideal circumstances. We used the model illustrated in Figure 1 (and described algebraically in the appendix) to measure sensitivity and predictive error of this indicator. We evaluated sensitivity and predictive error for the scenario of:

- perfect risk adjustment—ie, no variation in mortality rates among hospitals due to effects of casemix; and
- equal patient volume—ie, no variation in mortality rates among hospitals due to differences in mortality rate sampling variances.

We considered a hypothetical group of hospitals, each providing care to a sample of exactly  $N$  patients drawn randomly from a very large population of clinically similar patients. Hospitalized patients who receive good quality care have mortality risk of  $P_{D|GQ}$  whereas the  $P_{PQ}$  percent of the patients who receive poor quality care die at the higher rate of  $P_{D|PQ}$ . Of the hospitals,  $P_{PQH}$  percent are poor quality providers; in each of these hospitals, the percentage of patients who receive poor quality care is  $HPQR$  times greater than the percentage in good quality hospitals.

### Model Parameters

We derived values for the model's parameters from three sources: published research studies, publicly available hospital "report cards," and a data set of quality-of care review findings provided by the Texas Foundation for Medical Care (TFMC), the Medicare Peer Review Organization for Texas.

$P_{D|PQ}$ ,  $P_{D|GQ}$ , and  $PPQ$ . Few studies reported in the literature have attempted empirically to relate quality of care and patients' mortality risks. Two articles from the Rand Corporation's evaluation of the Medicare Prospective Payment System (PPS), however, provided precise estimates for such relations.<sup>10,11</sup> With a sample of 14,000 Medicare patients hospitalized before and after implementation of PPS, Kahn et al<sup>11</sup> utilized condition-specific explicit process criteria to assess the quality of care documented in patients' medical records. For a 1,200 case subsample of the five-condition sample, Rubenstein et al<sup>10</sup> used a sec-



and quality measurement methodology, called structured implicit review. Reviewers utilized a structured protocol of 27 questions covering such issues as the process of physician and nursing care, appropriateness of use of hospital services, patient prognosis, treatability of patient's condition, and overall assessment of quality of care during the hospitalization episode. This methodology was found to provide quality rankings generally similar to those developed with explicit review quality scales, but to yield a more sensitive indicator of quality as judged by mortality rate odds ratios.<sup>10,11</sup> For the subsample, 14.7% of the patients whose care was judged as good quality died within 30 days of admission compared with a 30-day mortality rate of 21.6% for patients who received poor quality care. For the post-PPS period, 12% of reviewed records involved poor quality care.<sup>†</sup> We utilized these estimates of  $P_{DIGQ}$ ,  $P_{DIPQ}$ , and  $P_{PQ}$  for our model.

**Number of Patients per Hospital (N).** For public reports on hospital mortality rate performance, patient samples for individual hospitals vary in size depending on hospital annual volume, number of months of data included, clinical condition(s) considered, and case exclusion criteria. Numbers of patients per hospital for a sample of mortality data reports are shown in Table 1. The relatively high medians for coronary artery bypass graft (CABG) surgery cases in New York and Pennsylvania reflect policies that limit performance of this procedure to small numbers of tertiary medical centers.<sup>12,13</sup> A large median sample size also is shown for the Michigan Hospital Association report that combines mortality data across six conditions: acute myocardial infarction (AMI), congestive heart failure (CHF), stroke, pneumonia, chronic obstructive pulmonary disease (COPD), and gastrointestinal bleeding.<sup>14</sup> For most conditions in Table 1, however, hospital samples are smaller, ranging from 133 to 367. For our modeling analyses, we initially used  $N = 200$ . Subsequently, we examined mortality statistics based on larger patient samples per hospital.

**$P_{POH}$  and HPQR.** We are unaware of any published estimate of the proportion of hospitals that

deliver unacceptably high levels of poor quality care [ $P_{POH}$ ], nor have we seen estimates for relative probabilities of patients receiving poor quality care in poor quality hospitals compared to good quality hospitals (HPQR). Quantifying these parameters requires quality-of-care evaluations for large samples of patients at large numbers of hospitals. To our knowledge, the only feasible sources for such data are the quality review data bases of Medicare Peer Review Organizations (PROs). From the Texas Foundation for Medical Care (TFMC), we were able to obtain records on 220,000 quality reviews performed during 1990 and 1991. We selected all cases admitted for AMI, CHF, pneumonia, or stroke in the 374 hospitals that had at least 50 admissions in these diagnoses. For each hospital, we determined the percentage of patients that experienced quality-of-care problems. We then sorted the hospital records in order of descending quality problem rates and arbitrarily chose the worst 10% for focus. In the worst 10% of Texas hospitals, the percentage of patients receiving poor quality care was 4.1 times greater than the average of the remaining 90% of hospitals. For our analyses, we let  $P_{POH} = 10\%$  and  $HPQR = 4.1$ .

## Results

### Expected Values for Sensitivity and Predictive Error

As shown in Table 2, with the parameters defined above, overall mortality rate across all hospitals was 15.5%. Among good quality hospitals, the mean (expected) mortality rate was 15.3%, and for poor quality hospitals it was 17.3%. The trim point (two-tailed test, 95% confidence limits) was 20.5%.

In our analysis, we have assumed that 10 out of every 100 hospitals are delivering poor quality care. As shown in Table 2, 1.12 of these 10 hospitals will be high outliers and thus correctly identified as poor quality providers—ie, true positives; however, 8.88 of the 10 will have mortality rates that fall below the trim point. They are false negatives, because they are not correctly identified as poor quality providers. Of the 90 good quality hospitals, 88.17 had mortality rates lower than the trim point and were correctly classified as good quality providers (true negatives). The 1.83 good quality hospitals with mortality rates above the trim point repre-

<sup>†</sup> $P_{PQ}$  during the Pre-PPS period was 25%.  $P_{DIGQ}$  is the weighted average of rates reported in Rubenstein et al<sup>10</sup> for good or very good quality care;  $P_{DIPQ}$  is the weighted average of rates reported for poor or very poor quality care (Rubenstein et al<sup>10</sup> Table 3, p 1977).

TABLE 1. Number Patients Per Hospital: Selected Mortality Data Reports

Conditions	Source and Reference	Number of Hospitals	Patients per Hospital		
			Minimum	Maximum	Median
CABG	PA HCCC*	41	65	960	403
CABG	NY DOH <sup>†</sup>	31	74	1393	530
AMI	CA OSHPD <sup>‡</sup>	395	1	917	133
AMI	PAHCCC <sup>§</sup>	49	30	381	133
AMI	MHA <sup>¶</sup>	116	30	769	115
Stroke	PA HCCC <sup>  </sup>	36	2	312	164
COPD	PA HCCC <sup>  </sup>	35	37	431	160
Pneumonia	PA HCCC <sup>  </sup>	36	35	364	145
CHF	PA HCCC <sup>  </sup>	35	95	607	367
6 Medical Dx <sup>#</sup>	MHA <sup>¶</sup>	157	41	4328	614

\*Pennsylvania Health Care Cost Containment Council (PHCCC). A consumer guide to coronary artery bypass graft surgery. Vol IV 1993 Data. PHCCC: Harrisburg, PA: 1995.

<sup>†</sup>New York State Department of Health (NYSDH). Coronary artery bypass surgery in New York State 1991-1993. NYSDH: Albany, NY: 1995.

<sup>‡</sup>Office of Statewide Health Planning and Development, State of California (OSHPD). Report of the California hospital outcomes project: hospital specific detailed statistical tables. OSHPD: Sacramento, CA: 1996.

<sup>§</sup>Pennsylvania Health Care Cost Containment Council (PHCCC). Focus on heart attack in Western Pennsylvania: A 1993 summary report for health benefits purchasers, health care provider, policy-makers, and consumers. Harrisburg, PA: PHCCC, 1996.

<sup>¶</sup>Michigan Hospital Association (MHA). Michigan Hospital Performance Report—May 1996. MHA: Lansing, MI: 1996.

<sup>||</sup>Pennsylvania Health Care Cost Containment Council (PHCCC) Hospital effectiveness report. Region 1. Reporting period: Jan 1-Dec 31, 1992. Harrisburg, PA: PHCCC, 1994.

<sup>#</sup>AMI, CHF, stroke, pneumonia, COPD, gastrointestinal bleeding.

sent false positives. In this case, sensitivity of the mortality rate trim point for identifying poor quality providers was 11.2%—it failed to identify 88.8% of the hospitals that are providing high levels of poor quality care. The predictive error, 62.1%, is the proportion of high outlier facilities that are good quality hospitals.

### Influence of Individual Parameters on Sensitivity and Predictive Error

The analysis in Figure 2 was based on best available estimates for the parameters of our model; however, actual values of these parameters may differ from those assumed. To determine how variation in each of the parameters might affect robustness of our conclusions about mortality rate identification of poor quality hospitals, we reevaluated the model multiple times with different values of each parameter, holding other parameters constant. Results are shown in Figures 3 through 6.

**Size of Patient Sample [N].** In Figure 3, we show sensitivity and predictive error of hospital mortality rate indicators as functions of number of patients per hospital. Because every one of the example reports listed in Table 1 included data on hospitals that served very small numbers of patients (minimums range from one to 95), we evaluated the accuracy of identifying poor quality hospitals that serve samples as small as 100 patients. With  $N = 100$ , sensitivity was 7% and predictive error was 71%. Both sensitivity and predictive error improved continuously as  $N$  increases, with sensitivity reaching 23% and predictive error decreasing to 41% with  $N = 600$ .

**Mortality Rate for Good Quality Care ( $P_{D:GQ}$ ).** To determine the degree to which accuracy of mortality rate identification of poor quality hospitals depends on the underlying mortality rate of diagnoses being considered, we examined levels of  $P_{D:GQ}$  ranging from 2.5% to 25.0%. Other model parameters were held constant, except  $P_{D:PQ}$  was varied to keep constant the ratio of

TABLE 2. Targeting Poor Quality Hospitals As High Mortality Rate Outliers, Based Upon Best Available Estimates for Model Parameters

Input Parameters		Targeting Results	
Mortality rate with GQ [P(O GQ)]	14.7%	Among 90 GQ hospitals and 10 PQ hospitals	
Mortality rate with PQ [P(O PQ)]	21.6%	True Positives	1.12
Patients in population receiving PQ [P(PQ)]	12.0%	False Negatives	8.88
Hospitals classified as PQ [P(PQH)]	10.0%	False Positives	1.83
HPQR	4.1	True Negatives	88.17
Number patients per hospital [N]	200		
Calculated parameters		Summary measures of targeting accuracy	
Patients at GQ hospital receiving PQ care	9.2%	Sensitivity	11.2%
Patients at PQ hospital receiving PQ care	37.6%	Predictive Error	62.1%
Expected mortality rate at GQ hospitals	15.3%		
Expected mortality rate at PQ hospitals	17.3%		
Expected overall patient mortality rate	15.5%		
Outlier trim point	20.5%		

HPQR. Hospital Poor Quality Ratio.

mortality risks for poor to good care ( $P_{DIPQ}/P_{DIGQ} = 1.47$ ).<sup>‡</sup> Results are shown in Figure 4. Both sensitivity and predictive error improved continuously with increasing levels of underlying mortality risk. For  $P_{DIGQ} = 2.5\%$ , sensitivity was 5.4% and predictive error was 79.1%. With the ratio  $P_{DIPQ}/P_{DIGQ}$  held constant at 1.47, if patients who receive good quality care die at a rate of 25.0%, then the mortality rate associated with poor quality care is 36.8%. Yet sensitivity remained less than 17%, and more than half of the hospitals identified as high outliers were actually good quality hospitals.

**Percentage of Patients Receiving Poor Quality Care [ $P_{PQ}$ ].** Rubenstein et al<sup>10</sup> reported that the proportion of Medicare patients receiving poor quality care in hospitals decreased from 25% in 1981 to 1982 to 12% in 1985 to 1986. Significant improvements in quality of care after implementation of the Medicare PPS also were noted by Kahn et al<sup>11</sup> based on explicit criteria process of care measurement. In Figure 5, we examine the implications of different levels of overall quality of

care for the accuracy of mortality rate identification of poor quality hospitals. As would be expected, at  $P_{PQ} = 2.5\%$ , sensitivity was low, less than 4%, and predictive error was greater than 85%. Both statistics improved with higher overall levels of poor quality in the population. If  $P_{PQ}$  reaches 25%, approximately double the proportion noted by Rand for the 1985–1986 period, predictive error decreases to 33%; ie, two thirds of the hospitals identified as poor quality providers on the basis of mortality rates were correctly identi-

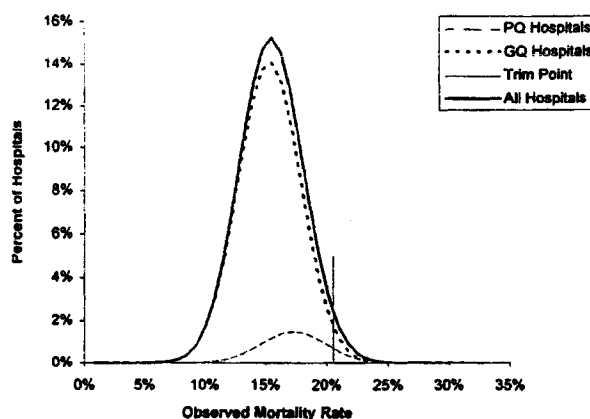


FIG. 2. Targeting poor quality hospitals as high mortality rate outliers, based upon best available estimates for model parameters.

<sup>‡</sup>For specific conditions examined by Kahn et al<sup>11</sup> using explicit criteria to measure quality of care, the ratios of mortality risk for poor quality care to good quality care was 1.52 for CHF, 1.33 for pneumonia, 1.33 for AMI, and 1.26 for stroke.

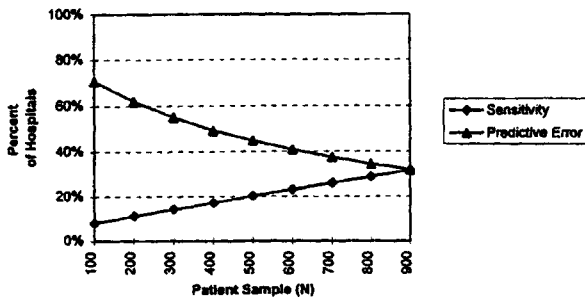


FIG. 3. Sensitivity and predictive error for different levels of N. Other model parameters as specified in Table 2.

fied. However, fewer than one third of the very poor quality hospitals were correctly identified as high outliers.

**Proportion of Poor Quality Hospitals [ $P_{PQH}$ ] and Hospital Poor Quality Ratio [HPQR].** For the analysis displayed in Figure 2, we assumed that 10% of hospitals were poor quality providers, and we found that only 11.2% of these providers were identified as high mortality rate outliers. If we were to focus on identifying the worst 5% of hospitals, or the very worst 2.5%, would we find mortality rate indicators to be more successful (sensitive) at identifying the poor quality providers?

Relations between  $P_{PQH}$  and sensitivity and predictive error are shown in Figure 6, where  $P_{PQH}$  ranges from 2.5% to 20%. In these analyses, HPQR was varied simultaneously because the average percentage of patients receiving poor quality care in the worst 5% of hospitals would be expected to be higher than in the worst 10%, and higher in the worst 2.5% than in the worst 5%.

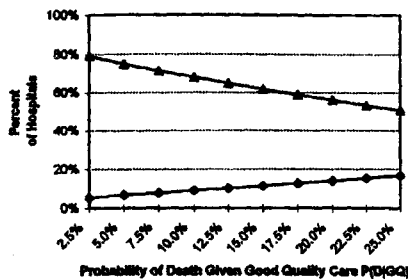


FIG. 4. Sensitivity and predictive error for different levels of P(D/GQ). Other model parameters as specified in Table 2, except mortality risk odds associated with poor quality held constant at 1.47.

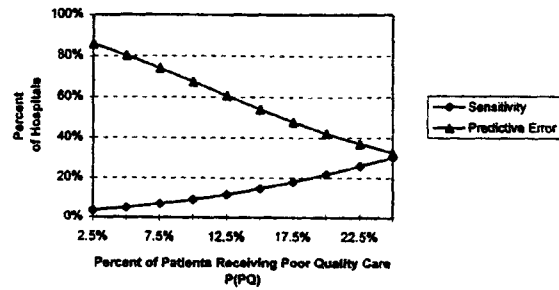


FIG. 5. Sensitivity and predictive error for different levels of P(PQ). Other model parameters as shown in Table 2.

ratio (HPQR) to be 6.6 for the worst 2.5% of hospitals. In these institutions the percentage of patients receiving poor quality care is 6.6 times greater than in the remaining 97.5% of hospitals. In the worst 5% of hospitals, the rate of poor quality care was 5.6 times higher than in the other 95% of hospitals. At higher values of  $P_{PQH}$ , however, the ratio remains relatively stable—HPQR = 4.1 with  $P_{PQH} = 10\%$ , HPQR = 4.4 at  $P_{PQH} = 15\%$ , and HPQR = 4.3 at  $P_{PQH} = 20\%$ . If poor quality care is heavily concentrated in only 2.5% of hospitals, the set of high mortality rate outliers would include 40% of the poor quality providers; however, 68% of outliers are actually good hospitals. Focusing on the worst 5% of hospitals, approximately 80% of the poor quality providers escaped detection as high outliers, and, of the high outlier facilities identified, nearly two thirds were actually good quality hospitals. How accurately do risk-adjusted mortality rates identify the worst 20% of hospitals? Identification was relatively good in terms of predictive error, because in this case only 50% of the high outlier hospitals were

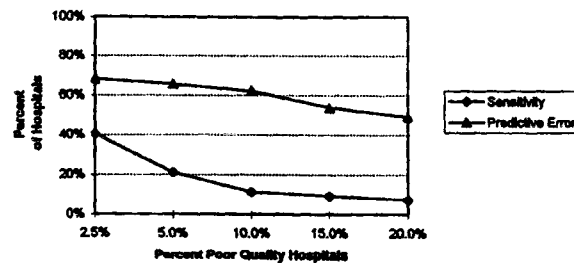


FIG. 6. Sensitivity and predictive error for different levels of P(PQH) and HPQR. Other parameters as in Table 2. For P(PQH) = 2.5%, HPQR = 6.6; for P(PQH) = 5%, HPQR = 5.6; for P(PQH) = 10%, HPQR = 4.1; for P(PQH) = 15%, HPQR = 4.4; and for P(PQH) = 20%, HPQR = 4.3.

false positives, but less than 8% of the poor quality hospitals were identified.

### Discussion

In this article, we have proposed a definition of accuracy for mortality rate indicators of hospital quality performance, and we have presented a model that allows precise determination of the potential accuracy of such measures. We found that under optimal conditions—perfect risk adjustment and no variation in patient volume among hospitals—we can expect that fewer than 12% of hospitals actually delivering poor quality care would be identified as high mortality rate outliers, and that of the facilities that are identified as outliers, more than 62% actually will be good quality providers. We examined how sensitivity and predictive error might vary with different estimates for model parameters, and under virtually all realistic assumptions for parameter values, sensitivity remained less than 20%, and predictive error was greater than 50%.

### Caveats and Limitations

The analyses presented here have limitations that could affect the generalizability of our reported conclusions. First, key parameter values utilized in this study were derived from published research on Medicare patients hospitalized for serious medical conditions. It is possible, therefore, that for other conditions and types of patients, the mortality risks associated with good quality care and poor quality care might differ from values used in our analyses.

Our use of 200 for the size of patient samples might be criticized as unfairly small. As demonstrated in Table 1, this value is not atypical for samples appearing in published hospital performance reports. Further, although accuracy of mortality-rate identification of poor quality hospitals improves as sample sizes increase, Figure 3 indicates that even with samples as large as 900 patients, a third of high-outlier hospitals were false positives.

Yet another possible criticism of our analysis is that when we investigated the implications of parameter values that differed from our base case assumptions, we varied only one parameter at a time. In a series of analyses not presented here, we also examined what we considered realistic variations in values for combinations of param-

eters. In general, ranges of resulting sensitivity and predictive error were not different from the extreme values observed when we varied parameters singly (Figs. 3–6).

There was one exception. Although we have no empirical estimates for  $P_{DIGO}$  and  $P_{DIPQ}$  for coronary artery bypass graft (CABG) surgery, it is not unreasonable to assume from data presented by Hannan et al<sup>15</sup> that values for these parameters are likely to be approximately 2% and 6%, respectively. From our Texas Foundation for Medical Care data base, we calculated for CABG surgery (DRGs 110 and 111) that patients in the worst performing 10% of hospitals were 5.4 times as likely to receive poor quality care as patients in the other 90% of hospitals. Using these parameter values, we found that with the median hospital volume ( $N = 530$ ) in the New York State CABG surgery mortality report referenced in Table 1, 80% of facilities identified as high mortality rate outliers were in fact poor quality hospitals. If  $N = 403$ , the median hospital volume shown in Table 1 for Pennsylvania's CABG mortality data report, predictive error was 25%. With the exceptions noted, other parameter values used in these analyses are as shown in Table 2.

### Alternatives to Mortality Rates As Indicators of Hospital Quality Performance

With the single exception of CABG surgery mortality, there can be little doubt that the currently available public data reports represent misinformation. The major force driving change in the United States health care system during the past decade has been employer efforts at value purchasing, and an essential element of value purchasing is access to information on vendor performance.<sup>16,17</sup> As employers increasingly limit available health plan options, they must be able to offer objective evidence that the managed care networks into which employees and dependents are being directed do not include low quality providers.

If risk-adjusted mortality rates are not suitable, what measures should be used for reporting on hospital performance? We believe that the best candidates are likely to be process-based indicators, such as degree of compliance with explicit, condition-specific criteria.<sup>18</sup> Process measures do not relate to the "bottom line" of medical care, as outcome measures are commonly thought to do. Evidence suggests, however, that process-based

measures are potentially more sensitive than outcomes for detecting quality of care differences among hospitals.<sup>19</sup> Moreover, hospitals find process-based measures, unlike outcome statistics, "actionable," ie, helpful in determining how quality of care can be improved.<sup>20</sup>

Nevertheless, to avoid dissemination of misleading information, as has been done and is being done with mortality rate measures, purchasers should demand that the accuracy of process-based indicators of quality be demonstrated before the indicators are used in public reports of hospital performance.

### References

1. **Park RE, Brook RH, Kosecoff J, Keesey J, Rubenstein L, Keeler E, et al.** Explaining variations in hospital death rates: Randomness, severity of illness, quality of care. *JAMA* 1990;264:484.
2. **Iezzoni LI.** The risks of risk adjustment. *JAMA* 1997;278:1600.
3. **Kazel R.** Hospital performance data under fire. *Crain's Detroit Business* April 14, 1997.
4. **Green J, Winfield N.** Report cards on cardiac surgeons: Assessing New York State's approach. *N Engl J Med* 1995;332:1229.
5. **Hofer TP, Hayward RA.** Identify poor-quality hospitals: Can hospital mortality rates detect quality problems for medical diagnoses? *Med Care* 1996;34:737.
6. **Zalkind DL.** Mortality rates as an indicator of hospital quality. *Hosp Health Serv Admin* 1997;42:3.
7. **Rutstein DD, Berenberg W, Chalmers TC, Child CG, Fishman AP, Perrin EB.** Measuring the quality of medical care: A clinical method. *N Engl J Med* 1976;294:582.
8. **Donabedian A.** The quality of medical care. *Science* 1978;200:856.
9. **McAuliffe WE.** A validation theory of quality assessment. In: Pena JJ, ed. *Hospital quality assurance: Risk management and program evaluation*. Gaithersburg, MD: Aspen Publishers, 1984.
10. **Rubenstein LV, Kahn KL, Reinisch EJ, Sherwood MJ, Rogers WH, Kamberg C, et al.** Changes in quality of care for five diseases measured by implicit review, 1981-1986. *JAMA* 1990;264:1974.
11. **Kahn KL, Rodgers WH, Rubenstein LV, Sherwood MJ, Reinisch EJ, Keeler EB, et al.** Measuring quality of care with explicit process criteria before and after implementation of the DRG-based Prospective Payment System. *JAMA* 1990;264:1969.
12. **New York State Department of Health (NYSDH).** Coronary artery bypass surgery in New York State 1991-1993. Albany, NY: NYSDH, 1995.
13. **Pennsylvania Health Care Cost Containment Council (PHCCC).** A consumer guide to coronary artery bypass graft surgery. Vol. IV 1993 Data. Harrisburg, PA: PHCCC, 1995.
14. **Michigan Hospital Association.** Michigan hospital performance report—May 1996. Lansing, MI: Michigan Hospital Association, 1996.
15. **Hannan EL, Kumar D, Racz M, Siu A, Chassin M.** New York State's cardiac surgery reporting system: Four years later. *Ann Thorac Surg* 1994;58:1852.
16. **Etheridge L, Jones SB, Lewin L.** What is driving system change? *Health Aff* 1996;15:11.
17. **Inglehart JK.** Competition and the pursuit of quality: A conversation with Walter McClure. *Health Aff* 1988;7:79.
18. **Huff ED.** Comprehensive reliability assessment and comparison of quality indicators and their components. *J Clin Epidemiol* 1997;50:1395.
19. **Mant J, Hicks N.** Detecting differences in quality of care: The sensitivity of measures of process and outcome in treating acute myocardial infarction. *BMJ* 1995;311:793.
20. **Thomas JW.** Report cards: Useful to whom and for what. *Jt Comm J Qual Improv* 1998;24:50.