FROM RESEARCH TO PRACTICE

How Severity Measures Rate Hospitalized Patients

John S. Hughes, MD, Lisa I. Iezzoni, MD, MSc, Jennifer Daley, MD, Linda Greenberg, MA

In the vigorous marketplace where managed care plans compete on the basis of cost and quality, doctors now expect that they will be judged, not only on clinical, but also on financial performance. Any credible ratings for doctors and hospitals that emerge from this process must adjust for baseline patient risk, because the sickest patients are not evenly distributed across providers. A variety of severity measures, particularly for hospitalized patients, are now available[1]; how well they work is crucial to the accuracy and fairness of performance assessments.

Whether evaluating colleagues or being judged themselves, physicians will bear most of the consequences of performance comparisons. Physicians therefore need to become informed about severity measures, not only for self-protection, but also to be able to use them appropriately and wisely. Already many physicians have been approached by their hospital administrators to help select a severity measure, either for internal quality management or to respond to external demands of payers and purchasers.

Understanding the many severity measures available for hospital patients is a challenge. Many are costly commercial products, and their underlying logic and rules for rating severity are protected as trade secrets.[2] Severity measures vary widely in the types of clinical and other data they employ and the ratings they produce. Even if the variables used for rating severity are revealed, the weights assigned to them and the exact scoring approach often are not. Without this information, it is virtually impossible for physicians to evaluate whether the measure reflects the way they really think about severity of illness.

From the Department of Medicine, West Haven Veterans Affairs Medical Center, West Haven, Conn. (JSH); Division of General Medicine and Primary Care, Department of Medicine, Harvard Medical School, Beth Israel Hospital, the Charles A. Dana Research Institute, and the Harvard-Thorndike Laboratory, Boston, Mass. (LII, JD, LG); Health Services Research and Development, Department of Medicine, Brockton/West Roxbury Veterans Affairs Medical Center, West Roxbury, Mass. (JD).

Supported by the Agency for Health Care Policy and Research, under grant RO1 HS06742-03. Dr. Daley is Senior Research Associate, Career Development Program of the Department of Veterans Affairs Health Services Research and Development Service. The views expressed are solely those of the authors.

Address correspondence to Dr. Hughes: Department of Medicine/111GIM, West Haven Veterans Affairs Medical Center, 950 Campbell Ave., West Haven, CT 06516.

In addition, the few objective comparisons of these severity measures published to date suggest that severity measures have very different statistical abilities to predict outcomes,[3-7] and sometimes produce very different ratings of patients[6,7] and hospitals.[3-5] Therefore, the choice of severity measure could substantially affect perceptions of provider performance.

This review introduces several of the issues involved in examining the various severity measures for hospitalized patients. We categorize the types of ratings severity measures produce and attempt to provide a sense of how they operate. Using several case studies, we demonstrate some of their differences and similarities, and examine the factors and strategies that drive their assessments of severity in specific circumstances.

CONCEPTUAL FRAMEWORK AND SCORING METHODS

We consider nine severity measures that are widely used to evaluate hospitalized patients across a range of diseases (Table 1).[8-17] Severity measures differ in important ways, including data sources, original purposes, definitions of severity, approaches to classification and scoring, and the time intervals during hospitalization from which data are extracted.[1]

Data Sources

The data source is the most important practical distinction among severity measures. We consider six "code-based" and three "chart-based" measures. Code-based measures require only the demographic data (age, gender) and International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis and procedure codes routinely available in existing computerized hospital discharge abstracts. In contrast, chart-based measures require clinical variables and laboratory data newly gathered from medical records, as well as demographic information and, sometimes, diagnosis and procedure codes. The disadvantage of code-based measures is that they cannot consider many useful clinical indicators of illness severity such as blood pressure, heart rate, or fever. Nevertheless, because they rely on existing computerized information, they are much less costly and hence more feasible for widespread, immediate use than chart-based measures.[1]

Table 1. Description of Severity Methods

System*	Definition of Severity/ Classification Approach	Comments
	Code-Based Measures	
Acuity Index Method (AIM)[1]	LOS within DRGs	In-hospital complications can contribute to scores; role in quality assessment may be limited.
1. Score	Integer scores 1–5	
2. LOS[‡]	LOS in days	
All patient Refined Diagnosis-Related Groups (APR-DRGs)[8]	Total hospital charges within adjacent DRGs[†] (classes 1–4)	In-hospital complications can contribute to scores; role in quality assessment may be limited.
Comorbidity Index[10,11]	Risk of death within 1 year of medical hospitalization (additive integer score)	Based on chronic comorbid conditions, not acute events; may not predict in-hospital outcomes as well.
Disease Staging[12]		In-hospital complications can contribute to scores; role in quality assessment may be limited.
1. Mortality probability	Probability of in-hospital death (0 to 1.0)	
2. LOS	LOS relative to all patients (avg. = 100)	
3. Resource Demand scale	Resource use relative to all patients (avg. = 100)	
Patient Mngmt. Categories (PMCs)		In-hospital complications can contribute to scores; role in quality assessment may be limited.
1. LOS	LOS (in days)	
2. Resource Intensity Score (RIS)[15]	Hospital resource use (average = 1.0)	
3. Severity Level (SL)[16]	In-hospital morbidity and mortaltiy (levels 1–7)	
Refined Diagnosis-Related Groups (R-DRGs)[17]	LOS, total hospital charges within adjacent DRGs (classes B,C,D for medical; classes A,B,C,D for surgical patients)	Early deaths (within 48 hours) are grouped separately. In-hospital complications can contribute to scores; role in quality assessment may be limited.
	Chart-Based Measures	
Acute Physiology and Chronic Health Evaluation II (APACHE II)[14]		Developed and validated for intensive care unit patients. Scores based on first 24 hours after admission.
1. Acute physiology score	In-hospital mortality for patients in intensive care units (integer score 0–60)	
2. Total APACHE score	Includes acute physiology score plus scores for age, major illness (integer score 0–71)	
Computerized Severity Index (CSI)[9]	Physiologic complexity, based on combinations of diagnoses and disease-specific signs and symptoms	Clinical data elements are specific to disease category. Both versions have admission and maximum scores.
1. Integer score	Integer scores ≥ 0	Sum of squared scores for individual clinical variables.
2. Severity level	Levels 1–4	
MedisGroups (Atlas MQ)		Both versions have admission and midstay scores. Same clinical data elements collected for all medical records.
1. Original version	Clinical instability indicated by in-hospital death (levels 0–4)	
2. Empirical version[13]	Probability of in-hospital death (0–1.0)	

**Citations are given in the References section.*
† "Adjacent DRGs" are formed by consolidating DRGs previously split by complications and comorbidities.
‡LOS = length of stay.

Software Versions

Six of the nine measures have different software versions that generate different types of ratings from the same data (Table 1). Patient Management Categories (PMCs), for example, generate three ratings: a predicted length of stay (LOS), a Resource Intensity Score (RIS), and a Severity Level (SL). Similarly, Disease Staging produces a mortality probability, a LOS scale, and a Resource Demand scale. The Computerized Severity Index (CSI) yields an integer score and a severity level on admission, as well as a maximum score and level based on the most abnormal values from throughout the admission.

Definitions of Severity

Definitions of severity vary considerably (Table 1) and include the probability of in-hospital mortality (Medis-

Groups, Disease Staging), the probability of death at 1 year (Comorbidity Index), the expected length of stay (Acuity Index Method [AIM], Disease Staging), the patient's physiologic complexity (CSI), and the expected resource use (All Patient Refined Diagnosis-Related Groups [APR-DRGs], PMCs RIS, Disease Staging Resource Demand scale). Because most are derived to predict a specific outcome, a severity measure's underlying definition of severity can limit its applicability to other outcomes.

Diagnostic Categories

Some methods generate severity scores within diagnostic categories. AIM scores, defined to predict LOS, are specific to the patient's Diagnosis-Related Group (DRG), so that the relation of each severity level to length of stay varies across DRGs. Similarly for APR-DRGs and Refined Diagnosis-Related Groups (R-DRGs), scores are assigned within "adjacent DRGs"—which are formed by combining two or more individual DRGs, each containing the same list of principle diagnosis codes, but differing based on whether certain comorbid conditions and complications are present as additional diagnoses. The developers of MedisGroups created their own system of 64 disease groups and performed logistic regressions within each to create probability-of-death scores; the variables selected by the regressions and their assigned weights differ across the 64 groups.

The remaining severity measures use one set of scores for all hospitalized patients. These scores have identical meaning regardless of the patient's disease. For example, a 0.05 likelihood of dying means the same thing for a patient with acute myocardial infarction as for one with colon cancer.

Scoring Approach

Each severity measure has its own scoring approach and range. Some array patients over an interval scale (e.g., probabilities of death ranging from 0 to 1.0), while others use ordinal scales (e.g., scores of 1, 2, 3, and 4). In interval scales, a one-unit difference has the same meaning along the full range of the scale. For example, a 0.1 increment in the probability of death is always a 0.1 increase, regardless of whether the probability is increasing from 0.2 to 0.3 or from 0.7 to 0.8. In contrast, one-unit increases may not have the same meaning along the full range of an ordinal scale—patients with scores of 4 may not be "twice as sick" as patients with scores of 2.

Among the severity measures we examine, ordinal scales take the form either of clinical stages, ranging in number from 3 (DRGs) to 7 (PMCs), or of integer scores, produced by summing or otherwise aggregating integer weights assigned to various clinical components or diagnoses, as in the Comorbidity Index and CSI integer scores.

The continuous interval scales produced by some severity measures also are of several forms, including probability of in-hospital death, or a prediction of LOS expressed in days (PMCs LOS, AIM LOS). Others calibrate a continuous index around an average LOS (Disease Staging) or the average resource consumption level for all hospitalized patients. Among the latter, the overall average is set at 1.00 for the PMCs RIS, and at 100.0 for Disease Staging Resource Demand scale. A score of 1.55 for the former, or 155.0 for the latter, for example, indicates 55% more resource consumption than for average inpatients.

Scoring Distribution

Severity measures with ordinal scales have different ranges and distributions of scores. For example, a score of 3 is the maximum possible for APR-DRGs, but only a modest score for the PMCs severity level, which has a maximum of 7. Furthermore, the proportion of patients assigned to each level of an ordinal scale may vary widely among severity measures.

Time Intervals

The time frame of the severity score, or the interval of the hospitalization from which data are gathered, also differs, depending largely on the data source. Code-based measures use discharge abstract information, which is based on the entire hospitalization—ICD-9-CM diagnosis codes represent all conditions treated during the hospital stay, regardless of when they occurred. Chart-based systems tend to rate patients based on information taken from a defined interval after admission. CSI admission ratings, for example, are based on clinical variables from the first 24 hours of hospitalization, or the first 12 hours for patients admitted to an intensive care unit. CSI also produces maximum ratings, based on variables from the entire hospital stay, including the admission time period. MedisGroups generates an admission score and probability of death from data obtained usually in the first 2 days, and "midstay" ratings generally covering days 3 through 7.

Time frame has important implications for the utility of a severity measure. Knowing what happened over the entire hospitalization is particularly important for predicting resource consumption (e.g., as for the APR-DRGs and R-DRGs). However, encompassing all conditions arising during the hospitalization compromises the utility of these methods for drawing inferences about hospital quality based, for example, on severity-adjusted death rates. Code-based measures cannot distinguish comorbid conditions present at the time of admission from conditions or complications arising after admission.[18,19] Complications may result from deficiencies in hospital care, and thus complication codes sometimes indicate problems with quality of care, not severity due to the patient's underlying disease.

Table 2. Illustrative Cases

Case	Summary
Case 1 Age: 75 years DRG: 89, Simple pneumonia and pleurisy, age >17, with cc* LOS: 10 days Total charges: $7,900 Discharge status: Alive Discharge diagnoses: 481 Pneumococcal pneumonia 038.2 Pneumococcal septicemia 288.8 White blood cell disease, not otherwise specified 413.9 Angina pectoris 285.9 Anemia V10.51 History of bladder malignancy Procedures: None	A 75-year-old woman with a history of angina, glaucoma, and previous bladder cancer was admitted with a 3-day history of nausea, cough, chills, and pleuritic pain. On exam, her temperature was 101° F, blood pressure 130/70, pulse 86, respirations 20. She was in no distress but had decreased breath sounds in the right upper lung fields. Chest X-ray showed right upper lobe consolidation. Arterial blood gas on room air: pH 7.50, pCO_2 29, pO_2 80. Leukocyte count was 22.5k, with 88 segmented forms, 5 bands. Sputum and blood cultures grew pneumococci. She was treated with penicillin, improved promptly, and went home on day 10.
Case 2 Age: 45 years DRG: 110, Other cardiothoracic procedures without pump LOS: 1 day Total charges: $17,310 Discharge status: Expired Discharge diagnoses: 410.91 Acute myocardial infarction (MI), initial treatment 414.9 Chronic ischemic heart disease, unspecified 785.51 Cardiogenic shock 135 Sarcoidosis Procedures: 37.2 Cardiac catheterization 37.61 Intra-aortic balloon pump	A 45-year-old woman with pulmonary sarcoidosis treated chronically with low-dose prednisone, but no cardiac history, presented to an outpatient facility complaining of chest pain and shortness of breath. Paramedics were summoned, who found the patient slumped in a chair without vital signs. Cardiopulmonary resuscitation was begun on the scene, but sinus rhythm was not restored until after several attempts at defibrillation in the emergency department. She was unresponsive to pain and hypotensive with evidence of poor peripheral perfusion after resuscitation, but her pupils were reactive. Electrocardiogram showed evidence of a large anterior wall MI. Prompt insertion of an intra-aortic balloon pump was followed by cardiac catheterization, which showed an occluded proximal left anterior descending artery. Multiple attempts at angioplasty and intracoronary thrombolysis failed, and she died 3 hours after admission.
Case 3 Age: 59 years DRG: 123, Acute MI, expired LOS: 11days Total charges: $17,710 Discharge status: Expired Discharge diagnoses: 410.71 Subendocardial infarction, initial treatment 250.10 Diabetic ketoacidosis, adult 276.5 Volume depletion 578.9 Gastrointestinal hemorrhage 427.5 Cardiac arrest Procedures: 37.23 Right/left heart cardiac catheterization 88.72 Echocardiogram 89.44 Cardiac stress test	A 59-year-old unmarried male bus driver with hypertension, poorly controlled type II diabetes, obesity, and angina was admitted with several hours of abdominal discomfort, belching, and dry heaves. He was diagnosed with a subendocardial MI based on cardiac enzymes (peak creatine kinase 225, 4% MB fraction) and lateral ST segment and T wave changes on electrocardiogram. Hospital course was uncomplicated initially except for poor control of diabetes and blood pressure. A submaximal exercise stress test on the sixth hospital day showed 1.5 mm of ST segment depression in inferolateral leads. Cardiac catheterization on the 10th hospital day showed normal left ventricular function, 70–90% stenosis of the circumflex artery, and a 50–70% stenosis of the proximal left anterior descending artery, and several noncritical stenoses of the right coronary artery. The following day he was found unresponsive with hypotension and bradycardia, and died despite resuscitation attempts. An autopsy showed a massive saddle pulmonary embolism.

(Continued)

CASE STUDIES

The five cases presented here illustrate several methodologic issues and differences among severity measures. The cases were selected by the authors from patients admitted to Beth Israel Hospital during 1991 and 1992 for rating by each of the severity measures. For each case, the authors recorded demographic data (age, gender), admission and discharge dates, total hospital charges, discharge disposition, and ICD-9-CM diagnosis and procedure codes taken directly from the chart face sheet prepared by the hospital's medical records department.

Table 2. Illustrative Cases (continued)

Case	Summary
Case 4 Age: 81 years DRG: 210, hip and femur procedures, with cc* LOS: 12 days Total charges: $20,900 Discharge status: Expired Discharge diagnoses: 820.21 Pertrochanteric femur fracture, closed, intertrochantic section 780.2 Syncope and collapse 401.9 Essential hypertension 599.0 Urinary tract infection 038.40 Gram-negative septicemia 518.5 Pulmonary insufficiency following trauma and surgery Procedures: 79.35 Open reduction of femur fracture with internal fixation	An 81-year-old widow, previously ambulatory and residing at a home for the aged, was admitted after she fell and broke her hip. She had a history of hypertension, a previous syncopal episode with a negative diagnostic workup, mitral regurgitation, and aortic regurgitation. Repair of her hip was delayed until day 4 by a urinary tract infection and a syncopal episode that occurred in the emergency department on admission. Her postoperative course was complicated by persistent fever, a pleural effusion, and a probable urinary tract infection. She was started on antibiotics on day 10, and was transferred to the intensive care unit on day 11 due to hypotension (systolic pressure 80), hypoxia, and possible sepsis. She was intubated and mechanically ventilated, but became more hypotensive despite intravenous dopamine and neosynephrine drips. She died on day 12.
Case 5 Age: 54 years DRG: 475, Respiratory system diagnosis with ventilator support LOS: 6 days Total charges: $13,800 Discharge status: Expired Discharge diagnoses: 482.1 Pneumonia due to *pseudomonas* 203.0 Multiple myeloma Procedures: 96.04 Endotracheal intubation	A 54-year-old woman with a 5-year history of multiple myeloma was admitted with dyspnea and diffuse bilateral infiltrates on chest x-ray. Her myeloma had first manifested with spinal cord compression, which was treated with radiation therapy and 1 year of chemotherapy. Six months before admission a recurrent paraspinal mass was resected but recurred after several weeks. She underwent another course of radiation therapy, and reduction and internal fixation of the left shoulder following pathologic fractures. She was bedbound and cared for by her husband at home, and was on no active treatment for myeloma at the time of admission. On exam her temperature was 98.6° F, pulse 120, respirations 36, blood pressure palpable at 92. Arterial blood gas on room air: pH 7.45, pCO_2 39, pO_2 63. Leukocytes were 8.0k with 87% segmented forms and no bands. Sputum grew *pseudomonas* and pleural fluid contained malignant cells. The patient initially requested maximal treatment, including intubation and mechanical ventilation. She was treated with intravenous antibiotics but deteriorated rapidly. On hospital day 3 she became hypoxic, was intubated, and was transferred to the intensive care unit with respiratory failure, necrotizing pneumonia, sepsis, and septic shock. Her respiratory failure worsened, and she become unresponsive on day 5. At the family's request, orders for do not resuscitate status and comfort measures only were written on day 5, and she died the next day.

**cc = comorbidity or complication.*

We also prepared day-by-day accounts of important clinical events and detailed lists of laboratory results, which we summarized in narrative format. Severity ratings were derived from the resulting abstracts, either by the vendors or, for Acute Physiology and Chronic Health II (APACHE II) and the Comorbidity Index, by ourselves. Discharge abstracts and brief narrative summaries for the five cases appear in Table 2.

We identified the major determinants of the individual severity ratings presented in Table 3, either based on the commentary provided by vendors on their ratings of the cases, or by reviewing the scoring logic presented in published materials. The vendors of two of the chart-based measures, MedisGroups and CSI, used the narrative summaries to score cases. For the other chart-based measure, APACHE II, the authors derived scores from the clinical and laboratory data in the narrative summaries. We obtained ratings for five of the six code-based measures from their vendors, who generated their ratings using only discharge abstract information, primarily the dis-

Table 3. Summary of Case Scoring for Illustrative Patients

	Range	Case 1	Case 2	Case 3	Case 4	Case 5
Ordinal scales						
AIM Score	1–5	4	2	4	3	1
APR-DRGs	1–4	4	4	4	3	2
CSI score						
Admission	1–4	2	4	1	2	3
Maximum	1–4	2	4	4	4	4
MedisGroups admission severity	0–4	2	2	2	1	3
PMCs Severity Level (SL)	1–7	6	5	5	7	4
Refined DRGs (R-DRGs)	D,C,B,A	B	A	B	A	B
Integer scores						
APACHE II						
Acute Physiology score	0–60	4	19	5	1	8
Total	0–71	10	26	8	7	10
Comorbidity Index	0–31	0	0	1	0	2
CSI integer score						
Admission	0+	25	65	20	29	35
Maximum	0+	25	65	73	83	80
Mortality prediction						
Disease Staging mortality	0–1.0	0.217	0.208	0.333	0.284	0.054
MedisGroups probability of death						
Admission	0–1.0	0.029	0.290	0.200	0.009	0.129
Midstay	0–1.0	0.029	—	0.009	0.013	0.415
Length of stay						
AIM LOS prediction (days)	>0	8.0	1.9	3.0	15.1	7.3
PMCs LOS prediction (days)	>0	12.7	12.1	11.2	21.3	8.4
Disease Staging LOS	Base = 100.0	158.5	40.7	65.4	215.7	149.3
Resource use						
Disease Staging Resource Demand	Base = 100.0	127.4	114.6	168.8	263.7	129.7
PMCs Resource Intensity Scale (RIS)	Base = 1.00	1.712	3.226	1.957	3.232	1.66

charge diagnosis and procedure codes. We scored the remaining code-based measure, the Comorbidity Index, ourselves, using the Deyo modification[10] of the Charlson index,[11] which employs only diagnosis codes.

Illustrative Case

Case 1 (Table 3), a 75-year-old woman admitted with typical, uncomplicated pneumonia, demonstrates the variety of output produced by severity measures. Among the ordinal scales, the two chart-based measures (CSI admission and maximum scores, MedisGroups) gave relatively low ratings, owing to the absence of severe clinical derangements on admission. In contrast, the four code-based measures all gave high ratings. The high PMCs rating (6 out of a possible 7), maximum APR-DRG score, and R-DRG class of B (the maximum for medical patients) all resulted from the ICD-9-CM code for septicemia. Two of the integer scores, APACHE II and CSI, based on clinical variables present on admission, were modestly elevated. The fact that the CSI admission and maximum scores were the same reflects the CSI's view that the patient did not deteriorate during the hospitalization. The other integer score, the Comorbidity Index, represents the burden of chronic comorbid illness rather than the extent of acute clinical complexity. Its rating of 0 reflected the absence of major coexisting illness.

The two mortality predictions diverged substantially. The code-based Disease Staging mortality prediction, like the code-based ordinal scales, was driven by the ICD-9-CM code for pneumococcal sepsis. In contrast, the MedisGroups probabilities of death, based on clinical findings such as low sodium, coma, high temperature, and high respiratory rate,[7] were much lower and did not change from admission to midstay.

For the LOS predictions, AIM and PMCs provided estimates differing by almost 5 days. The Disease Staging LOS index predicted a LOS 58.5% longer than the overall average for all hospitalized patients. The two measures of expected resource use also diverged: Disease Staging's Resource Demand rating was 27.4% higher than that of the average hospitalized patient, while the PMCs RIS predicted resource use was 71.2% higher than average.

Patterns of Differences Among Severity Measures

Code-Based Versus Chart-Based Measures and Time Windows

Cases 2 and 3, both patients with acute myocardial infarction, demonstrate how ratings can differ between code-based and chart-based measures. Both cases had relatively high PMCs severity levels of 5, as well as maximum R-DRGs and APR-DRGs scores. In case 2, the high rating for all three measures was due to the presence of cardiogenic shock on admission, while in case 3 high ratings resulted from the cardiac arrest following pulmonary embolism on the 11th hospital day. For case 2, the ratings fairly reflected the patient's condition on admission, but for case 3, they did not.

Case 4 also demonstrates the importance of the timing of data used for risk adjustment. The 81-year-old patient was fairly stable when admitted for hip fracture, as reflected by relatively low severity ratings from the chart-based ordinal scales (CSI score, MedisGroups admission severity). In contrast, code-based measures generally assigned higher ratings: PMCs severity level, APR-DRGs, and R-DRGs all gave maximum ratings, based on postoperative complications of Gram-negative septicemia and respiratory failure. The same pattern arose in the mortality prediction methods for this case, in which the code-based Disease Staging probability was substantially higher than MedisGroup's prediction.

For both cases 3 and 4, the inability of code-based measures to distinguish conditions present at admission from complications arising after admission limits their ability to detect quality problems. If the pulmonary embolism in case 3 or the postoperative complications in case 4 resulted from substandard quality, such as inadequate anticoagulation or poor postoperative monitoring, this causality could be obscured by the high severity ratings. Adjustments for the risk of in-hospital death or other adverse outcomes are most appropriately based on patient risk at the time of admission to the hospital; postadmission events and complications are examples of hindsight rather than prediction.[20]

The problem of discriminating comorbidities from complications applies to all code-based measures except the Comorbidity Index, which minimizes this problem by concentrating only on chronic comorbid conditions and ignoring most acute illnesses. This avoids confusing comorbidities with in-hospital complications, but also limits the ability of the Comorbidity Index to stratify acutely ill patients.

Implications of Diagnosis Coding

Code-based severity measures are susceptible to the vagaries of making diagnoses and coding them[21]: diagnosis codes may be omitted simply because of human error; physicians may not document conditions in the record, especially stable comorbidities, so coders cannot code them; clear-cut diagnostic criteria may not exist. Case 5 (pneumonia complicating end-stage multiple myeloma) illustrates the effects of coding inconsistencies. The original discharge abstract produced by the hospital omitted codes for septicemia and septic shock. The PMCs vendor indicted that with septicemia and shock codes, the severity level would have increased to the maximum of 7, the RIS would have risen from 1.66 to 3.011, and the projected LOS from 8.4 to 13.0 days. For the same case, the AIM score would have increased from 1 to 3 by adding the sepsis code, and the R-DRGs class would have risen to the maximum of B with coding of either septicemia or respiratory failure.

The chart-based measures may not be as susceptible to ICD-9-CM coding variation. Although the number of data elements required is much higher and gathering and processing of data are more complex, many chart-based data elements (e.g., laboratory test values) are often more easily identified and require less subjective interpretation than diagnosis coding.

Choice of Outcome Events

Assessments of the accuracy of severity rating may depend on the type of "risk," or outcome, being addressed. A severity measure that performs well in predicting in-hospital death may do less well predicting LOS or resource consumption and vice versa. In case 2, for example, the ordinal score produced by AIM was low, although the patient was seriously ill. If AIM were being used to predict risk of death, it would have performed poorly for case 2. If, on the other hand, we wished to predict LOS, as AIM intended, we would conclude that it had performed well—AIM predicted a very short LOS (1.9 days) for case 2.

The Comorbidity Index may be at a particular disadvantage for predicting short-term outcomes of hospitalization. Because the index counts only chronic conditions, it may give scores of only 0 or 1 to acutely severely ill patients (e.g., case 2) because they lacked major underlying chronic disease. The original purpose of the index was to predict survival 1 year after hospitalization. It may be inappropriate to ask it to stratify hospitalized patients for short-term outcomes.

Applying a severity measure to a use for which it was not originally designed, as happens frequently in actual practice, may or may not be a problem. Given that resource use, LOS, complexity of care, and the probability of death are frequently correlated, a measure that employs one definition of severity might be useful in predicting another. In case 2, however, the association broke down because the patient's severely ill state led to an early death and a short—and less expensive—hospital stay.

DISCUSSION

Severity measures for hospitalized patients differ in the sources and kinds of data used, time intervals of interest, definitions of severity, scoring algorithms, and type of ratings produced. As a result, different severity measures can yield different results when applied to the same group of patients. The interpretation of hospital performance may vary substantially as a result.

The most important distinction among severity measures is their source of data—whether they are code-based or chart-based. Code-based measures have limited usefulness for assessing quality of care, which often depends on distinguishing the patient's condition on admission from in-hospital complications. This limitation of code-based measures could be corrected by coding strategies to identify conditions arising during the hospitalization, such as adding another digit to existing ICD-9-CM diagnosis codes. Such a strategy would increase the complexity of coding discharge diagnoses, but would nonetheless be less costly than collecting clinical and laboratory data needed for chart-based measures.[22]

If the goal is to analyze costs rather than quality, on the other hand, code-based measures offer advantages. Considering all diagnoses treated during a hospitalization results in better predictions of costs and resource use. When predicting hospital costs, one would want to include complications, such as nosocomial infections, congestive heart failure, or cardiac arrest, even if there were evidence of poor-quality care—when complications occur, more resources are generally needed.

Key Points

Potential users of severity measures need to consider a number of issues[23,24]:

1. The choice of a severity measure will vary according to the context requiring it: whether one needs to address quality-of-care issues, to stratify costs, to evaluate new treatments, or to conduct research.
2. Severity measures are often used to predict outcomes different from those in their original operational definitions, and newer applications may not have been adequately evaluated. A severity measure that predicts one outcome well, such as probability of death, cannot necessarily be expected to perform as well for other outcomes, such as length of stay or resource use.
3. Severity measures produce several different types of ordinal scales and continuous interval scores. Furthermore, each severity measure distributes patients differently among its range of scores. The combination of differing outputs and varying distributions of scores means that comparisons among severity measures are difficult and that the methods cannot be used interchangeably. For example, an AIM level 3 means something different from a PMCs severity level 3, which also differs from an admission MedisGroups score of 3.
4. Finally, code-based and chart-based measures both have pros and cons. Chart-based approaches have more clinical content and are suitable for assessing severity at specified time periods, but they are expensive and sometimes cumbersome to apply. Code-based measures are considerably less expensive to implement, but sacrifice clinical content and may be less reproducible.

RECOMMENDATIONS

Physicians can take a number of specific steps when choosing among severity measures for hospitalized patients. They should first clearly identify how and why the measure is to be used. For example, an in-house review of care in the intensive care unit may require a measure with more clinical detail than an assessment of all patients for purposes of outside review. Physicians also need to evaluate whether their hospital's information systems can readily supply the data required by the severity measure. Furthermore, they should examine the practical implications of applying the various methods within their own practice settings. If the severity measure requires clinical laboratory data, problems could arise if the hospital lacks computerized laboratory reports or the ability to file results in patients' charts in a timely manner. Physicians can ask vendors for demonstrations of the logic that determines a severity measure's scores, and consider whether it is consistent with their own clinical reasoning. They should also determine whether the severity measure deals with unique characteristics of their own patient population. Does the severity measure have mechanisms to identify special populations, such as intravenous drug users or ethnic minorities, who may have special needs and higher rates of morbidity and mortality?

As an empirical test of how severity measures would perform in local settings, physicians could ask vendors to score severity for selected cases taken from the local hospital. Simultaneously, physicians should discuss the cases among themselves, making their own judgments about the patient's severity. By comparing their own assessments with scores from severity measures, physicians can evaluate which severity approach is most consistent with their own thinking.

The authors wish to acknowledge the substantial help of Betty Hotchkin in abstracting information from medical records and compiling severity ratings, and David Stone, PhD, and Robert Donaldson, MD, in reviewing earlier drafts of this article.

REFERENCES

1. Iezzoni LI, ed. Risk Adjustment for Measuring Health Care Outcomes. Ann Arbor, Mich: Health Administration Press; 1994.
2. Localio AR, Hamory BH, Sharp TJ, Weaver SL, TenHave TR, Lan-

dis JR. Comparing hospital mortality in adult patients with pneumonia. A case study of statistical methods in a managed care program. Ann Intern Med. 1995;122:125–32.
3. Iezzoni LI. "Black box" medical information systems: a technology needing assessment. JAMA. 1991;265:3006–7.
4. Thomas JW, Ashcraft MLF. Measuring severity of illness: six severity systems and their ability to explain cost variations. Inquiry. 1991;28:39–55.
5. 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.
6. Iezzoni LI, Ash A, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes. Ann Intern Med. 1995;28:763–70.
7. Iezzoni LI, Shwartz M, Ash AS, Mackiernan YD. Using severity measures to predict the likelihood of death for pneumonia in patients. J Gen Intern Med. 1996;11:23–31.
8. Edwards N, Honemann D, Burley D, Navarro M. Refinement of Medicare diagnosis related groups to incorporate a measure of severity. Health Care Fin Rev. 1994;16:45–64.
9. 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;29:305–17.
10. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol. 1992;45:613–19.
11. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chron Dis. 1987;40:373–83.
12. Gonnella JS, Hornbrook MC, Louis DZ. Staging of disease: a case-mix measurement. JAMA. 1984;251:637–44.
13. 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;30:128–41.
14. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985; 13:818–29.
15. Young WW, Swinkola RB, Zorn DM. The measurement of hospital case mix. Med Care. 1982;28:501–12.
16. Young WW, Kohler S, Kowalski J. PMC patient severity scale: derivation and validation. Health Serv Res. 1994;29:367–90.
17. Freeman JL, Fetter RB, Park H, et al. Diagnosis-related group refinement with diagnosis- and procedure-specific comorbidities and complications. Med Care. 1995;33:806–27.
18. Shapiro MF, Park RE, Keesey J, Brook RH. The effect of alternative case-mix adjustments on mortality differences between municipal and voluntary hospitals in New York City. Health Serv Res. 1994;29:95–112.
19. Hannan EL, Kilburn H, Lindsey ML, Lewis R. Clinical versus administrative data bases for CABG surgery. Does it matter? Med Care. 1992;30:892–907.
20. Blumberg MD. Risk adjusting health care outcomes: a methodologic review. Med Care Rev. 1986;43:351–93.
21. Hughes JS, Ash A. Reliability of risk-adjustment methods. In: Iezzoni LI, ed. Risk Adjustment for Measuring Health Care Outcomes. Ann Arbor, Mich: Health Administration Press; 1994; 263–86.
22. National Committee on Vital and Health Statistics, Subcommittee on Ambulatory and Hospital Care Statistics. Proposed Revision to the Uniform Hospital Discharge Data Set. Washington, DC: National Committee on Vital and Health Statistics; 1992.
23. Iezzoni LI. Case classification and quality of care. Issues to consider before making the investment. Qual Rev Bull. 1987;13:135–9.
24. Iezzoni LI. Examining the validity of severity measures in today's health policy context. J Gen Intern Med. 1995;10:406–7.