EDITORIALS

Examining the Validity of Severity Measures in Today's Health Policy Context

THE POLICY CONTEXT

Crafting a severity-adjustment strategy is an arcane methodologic pursuit, distant from daily medical practice. Nevertheless, findings derived from severity methods—risk-adjusted patient outcome measures—are increasingly used to evaluate clinical care, especially in competitive environments.[1, 2] Severity-adjusted patient outcomes are scrutinized by organizations ranging from state governments (e.g., in California, Colorado, Florida, Iowa, and Pennsylvania) to managed care companies to business coalitions.[3–6] Underlying this interest is the notion that providers must compete on "value," a melding of price and quality. In some parts of the country, "preferred providers" already are being selected using these performance measures to choose providers with the lowest cost—*levels of "quality" ostensibly being equal.* Thus, given their potential impact on clinical practice, the content and meaningfulness of performance measures should be of direct concern to physicians.

Despite its appeal, measuring provider value is difficult. Evaluating cost is complicated, and judging quality is even more problematic. Few valid quality measures are available, and existing health care databases rarely permit insight into important patient outcomes, such as functional status or quality of life. In most settings, the only data widely available are those produced as a by-product of billing or administration. Despite the limited clinical content of these data, they almost uniformly indicate whether patients lived or died. Because of its ready availability, information about death has thus become a staple of outcomes assessment, despite debate about its relation to provider quality.

Most efforts to compare patient death rates across providers recognize the importance of controlling for illness severity[7, 8]—some providers' patients are sicker (e.g., at higher risk of imminent death) than others. While few argue with this premise, how best to measure severity is unclear. A variety of commercial severity measures are now marketed to hospitals, states, governments, and even business leaders.[3–6] Few independent investigators have evaluated these measures.

One of the most successful proprietary severity measures is MedisGroups, marketed by MediQual Systems, Inc., and mandated for use in all Pennsylvania hospitals and all except small facilities in Colorado. Public release of MedisGroups severity-adjusted mortality rates for Pennsylvania hospitals has had an important impact on health care delivery in the state. MedisGroups severity-adjusted data were used by one large employer in the central part of the state to discourage its employees from seeking care at the local university medical center. Suburban facilities employ these data to lure patients from higher-priced, inner-city centers. Some of this information even attracted national attention, especially the comparison of provider death rates and average charges for coronary artery bypass graft (CABG) surgery.[9] President Bill Clinton cited findings from that report in his September 22, 1993, health care reform address to a joint session of Congress:

> We have evidence that more efficient delivery of health care doesn't decrease quality. . . . Pennsylvania discovered that patients who were charged $21,000 for [CABG] 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.[10]

QUESTIONS OF VALIDITY

Given the impact of MedisGroups-derived data, this method should itself undergo independent and objective examination, as in the research described by Fine et al. in this issue of the *Journal.*[11] What constitutes sufficient scrutiny of a severity method in the current health policy context? One answer involves determining its "validity." But as Donabedian observed:

> . . . The concept of validity is itself made up of many parts. . . . The question of validity 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.[12]

The work by Fine and colleagues addressed one aspect of the first domain—the precision of the models; unfortunately, they were unable to return to the original medical records to determine the accuracy of the MedisGroups data elements themselves. The authors also comment on the implications of their work for drawing inferences about hospitals and patients. These various

issues warrant further examination before one can claim that MedisGroups or any other severity method, such as the pneumonia severity of illness index (PSI) of Fine, is valid for specific uses.

MODEL PRECISION

The PSI was developed using appropriate statistical techniques and intelligent clinical input. Its clinical credibility is one of its most attractive features, and contributes to an overall sense of face validity—the PSI, "on the face of it," appears to measure what it claims to measure. MedisGroups' developers recognized the methodologic and clinical appeal of empirically derived, disease-specific models, and have replaced the version of MedisGroups evaluated by Fine with new measures for 64 disease groups, including pneumonia.[13] The empirical MedisGroups pneumonia model also has good clinical face validity; it is founded on 19 clinical variables, including respiratory rate, systolic blood pressure, temperature, arterial pH, percentage bands on white blood cell count, chest radiograph findings, and histories of cancer and immunocompromised state.[14] In testing by independent investigators using a comparable database of pneumonia patients,[15] the new MedisGroups produced a cross-validated C statistic (equal to the area under a receiver operating characteristic curve[16]) of 0.85, similar to that of the PSI.

The true test of the precision of any model, such as the PSI or the empirical MedisGroups, is its application to an entirely different data set from that used in model development. This remains to be done for both methods.

VALIDITY OF INFERENCES ABOUT HOSPITALS

It is not surprising that the PSI and the original MedisGroups score flagged different hospitals as good and bad mortality outliers—the two severity methods appear to be measuring slightly different things. One study found that even the original and new empirical MedisGroups methods disagreed on the mortality outlier status of eight of 105 hospitals.[15]

Although statistical performance measures for the PSI (and new, empirical MedisGroups) are reasonably good, neither explains close to 100% of the variation in patient outcomes. The presumption underlying the use of severity-adjusted death rates as hospital performance measures is that some of the difference between observed and expected hospital death rates can be explained by quality differences. That hypothesis is as yet unproven, and could not be assessed using the study design of Fine. The literature addressing the relationship of hospital mortality rates and quality of care yields inconsistent conclusions. Several reports link higher-than-expected mortality rates to substandard care,[17–19] while some do not,[20] and others provide equivocal conclusions.[21–24] Factors not captured in the MedisGroups Comparative Database could explain some of the differences across facilities, such as do-not-resuscitate (DNR) practices. For example, one study found statistically significant variations in rates of DNR use across 13 intensive care units, unexplained by patient age, prior health status, diagnosis, or severity of illness.[25]

Reason suggests that before using a measure to evaluate provider quality, the measure itself should be scrutinized to prove that it does indeed measure quality. In the current health policy context, however, the rules of evidence and proof appear reversed. Because it is often the only measure available, organizations will continue using severity-adjusted mortality rates as an indicator of quality until someone proves, definitively, that it is not. It is unlikely that this definitive study will be conducted any time soon: the research is expensive and requires defining a "gold standard" quality measure. Nevertheless, the results of Fine[11] and others[15] should suggest, at a minimum, that severity-adjusted hospital death rates be interpreted cautiously. Judgments about hospitals may vary using different severity measures.

VALIDITY OF INFERENCES ABOUT PATIENTS

Finally, there is an even-more controversial use of severity methods—to direct individual patient care. This application is suggested by Fine et al. in the last paragraph of their abstract: the "PSI's ability to accurately identify patients at extremely low risk of death supports its use in the identification of patients with pneumonia who may be treated effectively and safely with less intensive forms of therapy."[11] This use is attractive in an era of managed care and constrained resources, but the research was not designed to test this use, and advocating it is premature. Knowing that a particular group of patients who were given hospital care had low death rates is not the same as knowing that this group of patients would do well without such treatment.

Most importantly, the PSI can reliably identify a class of patients at low risk of dying but it might not identify the particular patient in that class who has a high risk of dying. One example is the pneumonia patient with debilitating chronic illness who desires "comfort measures only." In this situation, standard blood tests on which severity assessments are based (e.g., serum sodium, glucose, blood urea nitrogen, hematocrit, arterial pH) will not be performed. Without testing, no acute clinical derangements will be identified, and both the MedisGroups score and the PSI score will suggest a low risk of death. In this circumstance, however, the low severity rating does not represent the absence of severe disease.—**Lisa I. Iezzoni, MD, MSc,** *Associate Professor of Medicine, Division of General Medicine and Primary Care, Department of Medicine, Harvard Medical School, Beth Israel Hospital, the Charles A. Dana Research Institute, and the Harvard–Thorndike Laboratory, Boston, MA 02215*

REFERENCES

1. Epstein AM. Changes in the delivery of care under comprehensive health care reform. N Engl J Med. 1993;329:1672–6.
2. Jost TS. Health system reform. Forward or backward with quality oversight. JAMA. 1994;271:1508–11.
3. United States General Accounting Office. Health Care Reform. "Report Cards" Are Useful but Significant Issues Need to Be Addressed. (GAO/HEHS-94-219). Washington, DC: United States General Accounting Office; Health, Education, and Human Services Division, 1994.
4. United States General Accounting Office. Employers Urge Hospitals to Battle Costs Using Performance Data. (GAO/HEHS-95-1). Washington, DC: United States General Accounting Office; Health, Education, and Human Services Division, 1994.
5. Iezzoni, LI, Shwartz M, Restuccia J. The role of severity information in health policy debates: a survey of state and regional concerns. Inquiry. 1991;28:117–28.
6. Iezzoni LI, Greenberg LG. Widespread assessment of risk-adjusted outcomes: lessons from local initiatives. Joint Commission J Qual Improv. 1994;20:305–16.
7. Selker HP. Systems for comparing actual and predicted morality rates: characteristics to promote cooperation in improving hospital care. Ann Intern Med. 1993;118:820–2.
8. Kassirer JP. The use and abuse of practice profiles. N Engl J Med. 1994;330:634–6.
9. Pennsylvania Health Care Cost Containment Council. Coronary Artery Bypass Graft Surgery. Technical Report. Harrisburg, PA: Pennsylvania Health Care Cost Containment Council, 1992.
10. The White House Domestic Policy Council. Health Security: The President's Report to the American People. Washington, DC: The White House Domestic Policy Council, 1993;100.
11. Fine MJ, Hanusa BH, Lave JR, et al. Comparison of a disease-specific and a generic severity of illness measure for patients with community-acquired pneumonia. J Gen Intern Med. 1995;10:359–68.
12. Donabedian A. Explorations in Quality Assessment and Monitoring. Volume I. The Definition of Quality and Approaches to Its Assessment. Ann Arbor, MI: Health Administration Press, 1980;101.
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. Iezzoni LI. Dimensions of risk. In: Iezzoni LI (ed). Risk Adjustment for Measuring Health Care Outcomes. Ann Arbor, MI: Health Administration Press, 1994;112.
15. Iezzoni LI, Shwartz M, Ash AS, Hughes JS, Daley J, Mackiernan YD. Severity measurement methods and predicting pneumonia deaths. Med Care. In press, 1995.
16. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med. 1984;143–52.
17. Keeler EB, Rubenstein LV, Kahn KL, et al. Hospital characteristics and quality of care. JAMA. 1992;268:1709–14.
18. Kuhn EM, Hartz AJ, Gottlieb MS, Rimm AA. The relationship of hospital characteristics and the results of peer review in six large states. Med Care. 1991;29:1028–38.
19. Hartz AJ, Kuhn EM, Green R, Rimm AA. The use of risk-adjusted complication rates to compare hospitals performing coronary artery bypass surgery or angioplasty. Int J Technol Assess Health Care. 1992;8:524–38.
20. Park RE, Brook RH, Kosecoff J, et al. Explaining variation in hospital death rates, randomness, severity of illness, quality of care. JAMA. 1990;264:484–90.
21. Dubois RW, Rogers WH, Moxley JH, Draper D, Brook RH. Hospital inpatient mortality. Is it a predictor of quality? N Engl J Med. 1987; 317:1674–80.
22. Dubois RW, Brook RH. Preventable deaths: who, how often, and why? Ann Intern Med. 1988;109:582–9.
23. Fink A, Yano EM, Brook RH. The condition of the literature on differences in hospital mortality. Med Care. 1989;27:315–36.
24. Thomas JW, Holloway JJ, Guire KE. Validating risk-adjusted mortality as an indicator for quality of care. Inquiry. 1993;30:6–22.
25. Zimmerman JE, Knaus WA, Sharpe SM, Anderson AM, Draper EA, Wagner DP. The use and implications of do not resuscitate orders in intensive care units. JAMA. 1986;255:351–6.