Interpreting Measurement Data for Quality Improvement: Standards, Means, Norms, and Benchmarks

Richard C. Hermann, M.D., M.S.
Scott Provost, M.M., M.S.W.

Performance measurement has been proposed as a means of identifying and addressing quality problems, establishing accountability between payers and providers, informing oversight activities, and selecting plans and providers. Many delivery systems, government agencies, payers, and oversight groups have begun to implement measurement-based quality assessment to encourage improvements in mental health care (1,2). Interpreting results from these activities and achieving best practices have been hindered by the scarcity of comparative results.

Results for some measures—for example, rates of inpatient restraint—are meaningful only in comparison with results from facilities that treat similar populations. Other measures are based on a directional scale—for example, a low rate indicates poor conformance to a guideline, while a high rate indicates excellent conformance. However, results from these measures can still be difficult to interpret, because high conformance rates may not be feasible in difficult-to-treat populations.

This column describes several types of data that are useful for interpreting quality-measurement results: standards, means, norms, and benchmarks. Each can be useful in measuring quality, identifying best practices, and improving care. We use a narrow definition of a best practice, adapted from Donabedian's framework of structures, processes, and outcomes of health care (3). In this construct, a "practice" is a process—the means by which a delivery system interacts with and influences patients. A "best practice" is a process that has been demonstrated to produce superior results, such as better patient outcomes.

To illustrate a challenge confronting clinicians and managers in the course of quality-improvement work, we describe a common situation: A community mental health center named CMHC, Inc., decides to adopt a report card of performance measures to assess quality of care. The center selects measures that reflect the needs of its patient population and staff concerns about where care might fall short. One concern is that many of the patients may not be completing prescribed courses of antidepressants. Thus CMHC's report card includes a measure to assess adherence after initiation of medication treatment for depression. The measure assesses the proportion of patients with major depressive disorder who complete a 12-week acute phase of antidepressant treatment. Data from CMHC are collected and analyzed, and clinicians and managers meet to review the results. They learn that 55 percent of patients initiating a new course of antidepressant treatment completed the acute phase. How should they interpret this result?

Clinically, the result seems meaningful. Among patients sufficiently depressed to have started an antidepressant, nearly half did not complete a course adequate for remission to be achieved and sustained. The significance of the results is less clear in terms of CMHC's quality of care. Perfect performance on the measure is not a realistic standard. Some patients choose to discontinue antidepressants; their decision may reflect their preference or clinical complexity rather than poor-quality care. Yet lower rates may identify opportunities for improvement; research suggests that clinicians can influence compliance by educating patients, providing close follow-up, treating adverse effects, and addressing nonresponse (4). Further information would be useful for interpreting CMHC's performance. What performance would represent an excellent level of care? What constitutes average care? What results would indicate a clear need for improvement?

Several types of data can provide a starting point for addressing these questions, including standards, means, norms, and benchmarks. Some of these terms have been used inconsistently, although recent years have seen convergence toward more precise definitions. Standards are numerical performance expectations established by individuals or groups. For example, the Veterans Health Administration's (VHA's) office for quality and performance established a standard for rates of screening for depression in primary care. To be considered successful in meeting VHA's mandate, medical centers and clinics must screen 87 percent of primary care patients (5).

Standards may be based on statistically derived thresholds, reflect expert consensus, or be set arbitrarily. Because standards can be established without data, they are one of the most

The authors are affiliated with the Center for Quality Assessment and Improvement in Mental Health, Harvard Medical School and School of Public Health, McLean Hospital, 115 Mill Street, Belmont, Massachusetts 02478 (e-mail, richard_hermann@cqaimh.org). William M. Glazer, M.D., is editor of this column.

common sources of comparative information. A national inventory of quality measures for mental health care (available at www.cqaimh.org/ quality.html) found that 16 percent of more than 300 measures were accompanied by performance standards (6).

Published results from research studies or quality-assessment initiatives—usually available as means or averages—can provide a basis for preliminary comparisons. Published results should be used cautiously, because they may reflect unique characteristics of the population studied and the quality of care received. Their use is illustrated in our continuing case study. Report card results showed that 55 percent of CMHC's patients completed the acute phase of treatment. The next step was to determine whether this result represented a high quality of care, or an opportunity for quality improvement. Published results from use of the measure revealed that acute-phase completion rates ranged from 19 percent in a statewide sample of Medicaid beneficiaries to 59 percent in a nationwide sample of commercial health plan enrollees (3,7,8). The center's performance on the high end of the range suggested that the results were satisfactory.

Results are available for an increasing number of measures. A report recently prepared for the Substance Abuse and Mental Health Services Administration summarized published results from the application of more than 50 quality measures for mental health and substance abuse care (9). In comparing results obtained locally with results from published reports, several questions should be considered. What is the sample size? Results from studies of quality of care may be drawn from administrative data for tens of thousands of patients across hundreds of facilities, or drawn from medical records for a few dozen patients on a single inpatient unit. Smaller samples are less likely to be representative of the broader population. Did the study use a convenience cohort, or was a sampling strategy used? Norms—average results for large, representative population-based samples—can provide a useful reference point from which to compare results for large, diverse populations, but few have been developed for mental health quality measures.

One of the most important questions to consider is how characteristics of patients sampled in published studies compare with characteristics of samples studied locally. Patient characteristics may influence results for reasons unrelated to quality of care. For example, more severe depression, comorbid substance abuse, or less social support might all contribute to a lower likelihood of completion of the acute phase of antidepressant treatment.

More accurate comparisons of quality can be attained through case-mix adjustment, a statistical process of accounting for differences among patients' clinical and demographic characteristics when assessing processes or outcomes of health care. Also known as risk adjustment, case-mix adjustment may be needed to compare providers fairly in terms of their performance on quality measures (10). Stratification is a simple form of case-mix adjustment. Patients can be compared within categories—or strata—defined by one or more characteristics, such as the presence of a comorbid condition. This approach allows for comparison among more homogeneous populations. Multivariate modeling is a more sophisticated form of risk adjustment that allows for simultaneous adjustment for numerous characteristics. Not all quality measures require adjustment; some focus only on providers' behavior and are not influenced by patient characteristics (2). However, the development of case-mix adjustment models has lagged behind the development and use of quality measures. Only 13 percent of measures in a national inventory included methods for case-mix adjustment (6).

An additional shortcoming of average results relates to the goals of a quality-improvement initiative. Improving a problem area to an average performance level might be an important preliminary goal, but ultimately systems of care should aspire to excellence. Norms and averages provide little guidance on where excellence lies on the performance continuum. Benchmarking, a statistical process borrowed from industrial models of quality improvement, can provide numerical goals reflecting excellent yet achievable performance. Benchmarks reflect results of the highest-performing organizations in an industry (11). These results can be used by other organizations to interpret their own performance and develop numerical standards for quality improvement. Organizations that use industry averages as performance goals may be reinforcing a status quo, whereas benchmarks can help an organization to aim for the best possible result.

Weissman, Kiefe, and their colleagues (12) have pioneered the application of statistical benchmarking to health care. They define benchmarks operationally as the performance achieved by the top 10 percent of providers in a sample, adjusted for the number of patients per provider. This approach is objective, reproducible, and accounts for small sample sizes that can otherwise skew results.

Benchmarking does not obviate the need for case-mix adjustment. On the contrary: for clinicians to accept another organization's performance as relevant, they must be convinced that the results were obtained in a comparable patient population or that statistical analyses sufficiently adjusted for differences. Nevertheless, benchmarking holds considerable potential for motivating improvements in care. In a randomized controlled study, Kiefe and colleagues (13) compared two groups of primary care physicians undertaking a standardized quality-improvement protocol. Both groups received periodic feedback on their individual performance rates, but the group that received data on their individual performance in comparison with a statistical benchmark achieved significantly greater improvement.

The broader concept of benchmarks has been usefully applied in mental health care (14,15). But statistical benchmarks reflecting high levels of achievable performance are only now being developed (9). Our case example illustrates the potential of these benchmarks to enhance quality improvement activities. Although averages from published reports placed CMHC at the high end of performance on the antidepressant treatment measure, a statistical benchmark suggested a different interpretation. The center's 55 percent conformance rate

lagged well behind the benchmark of 91 percent (9). In response, CMHC clinicians began to discuss methods for improving treatment adherence.

Once an opportunity for improvement is identified, several sources of information about possible interventions are available. Each person involved in treatment may have insight into why the system produces certain results and what aspects might be improved. Best practices can come from research studies, disease management models, and facilities that are already achieving excellent results.

The center's inquiry yielded several promising practices. On the basis of research studies, CMHC staff began calling patients after initiation of or change in their medication regimen to answer questions or address adverse effects. In addition, clinicians made a site visit to an outpatient mental health clinic that was among the region's top performers on the measure. They adopted an innovation the clinic developed—a "best practice" in that it enabled the clinic to improve its performance. Patients receiving an initial antidepressant were scheduled to attend an educational session on medication treatment for depression and reasonable expectations for response.

Over the next six months, while implementing these changes in practice, CMHC continued to measure adherence on a monthly basis. The results informed their decisions about which interventions to continue and whether further changes were needed.

Measurement-based quality assessment and improvement is at an early stage of development in behavioral health care. We are just beginning to see the emergence of a supporting infrastructure: standardized data elements, common measures and specifications, and improvement strategies of proven effectiveness (2). Sources of comparative data are a crucial part of this foundation. Table 1 summarizes strengths and limitations of each metric. Benchmarks and other types of comparative data provide a means of identifying quality improvement opportunities. "Best practices" describe changes in practice that can contribute to the achievement of desired results. ♦

Table 1

Types of comparative data for interpretation of quality-measurement results

Data type	Strengths	Limitations
Standards	Set numerical expectations for performance Can reflect expert judgment Do not require extensive data	Can lack empirical foundation Can set expectations that are either unrealistic or too easily achieved
Means	Available from previous applications of measures Provide a basis for preliminary comparisons Multiple samples may increase utility	May reflect unique characteristics of the populations studied Case-mix adjustment is often lacking Comparison with average results does not encourage excellence
Norms	Provide a useful reference point for assessing treatment for large, diverse populations Can be developed for specific subpopulations	Available for few measures Comparison with average results does not encourage excellence
Statistical benchmarks	Represent excellent yet achievable care Derived by using objective, reproducible methods	Limited application to quality measures in mental health Stratification or other form of case-mix adjustment may still be needed

Acknowledgments

This work was supported by grant K08-MH001477 from the National Institute of Mental Health and by grant R01-HS10303 from the Agency for Health Care Research and Quality, and by the Substance Abuse and Mental Health Services Administration.

References

1. Hermann R, Regner J, Yang D, et al: Developing a quality management system for behavioral healthcare: the Cambridge Health Alliance experience. Harvard Review of Psychiatry 8:251–260, 2000
2. Hermann R, Palmer R: Common ground: a framework for selecting core quality measures. Psychiatric Services 53:281–287, 2002
3. Donabedian A: Exploration in Quality Assessment and Monitoring: The Definition of Quality and Approaches to Its Assessment. Ann Arbor, Mich, Health Administration Press, 1980
4. Chen A: Noncompliance in community psychiatry: a review of clinical interventions. Hospital and Community Psychiatry 42:282–287, 1991
5. Veterans Health Administration Office of Quality and Performance: FY2002 VHA Performance Measurement System: Technical Manual. Washington, DC, Nov 8, 2001 (updated Mar 8, 2002)
6. Hermann R, Leff H, Palmer R, et al: Quality measures for mental health care: results from a national inventory. Medical Care Research and Review 57(suppl 2):135–153, 2000
7. Melfi C, Chawla A, Croghan T, et al: The effects of adherence to antidepressant treatment guidelines on relapse and recurrence of depression. Archives of General Psychiatry 55:1128–1132, 1998
8. Kerr E, McGlynn E, Van Vorst K, et al: Measuring antidepressant prescribing practice in a health care system using administrative data: implications for quality measurement and improvement. Joint Commission Journal on Quality Improvement 26:203–216, 2000
9. Hermann R, Chan J, Chiu W, et al: Interpreting Findings From Quality Measurement Initiatives in Mental Health and Substance Abuse: Use of Published Data and Statistical Benchmarks. Report for the US Substance Abuse and Mental Health Services Administration, Center for Quality Assessment and Improvement in Mental Health, 2003. Available at www.cqaimh.org
10. Hermann R: Risk adjustment for mental health care, in Risk Adjustment for Measuring Health Care Outcomes. Edited by Iezzoni L. Ann Arbor, Mich, Health Administration Press, in press
11. Anderson CA, Cassidy B, Rivenburgh P: Implementing continuous quality improvement (CQI) in hospitals: lessons learned from the International Quality Study. Quality Assurance in Health Care 3:141–146, 1991
12. Weissman N, Allison J, Kiefe C, et al: Achievable benchmarks of care: the ABCs of benchmarking. Journal of Evaluation in Clinical Practice 5:269–281, 1999
13. Kiefe C, Allison J, Williams O, et al: Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. JAMA 285:2871–2879, 2001
14. Macias C, Barreira P, Alden M, et al: The ICCD benchmarks for clubhouses: a practical approach to quality improvement in psychiatric rehabilitation. Psychiatric Services 52:207–213, 2001
15. Mojtabai R, Lavelle J, Gibson PJ, et al: Gaps in use of antipsychotics after discharge by first-admission patients with schizophrenia, 1989 to 1996. Psychiatric Services 53:337–339, 2002