

# The Performance of Performance Measures

Karen S. Kmetik, PhD; Jeanette Chung, PhD; and Shannon Sims, MD, PhD

**A**ndrew Auerbach, MD, MPH, and colleagues recently argued that *interventions* (emphasis added) to improve the quality of healthcare should meet the same standards of evidence that are applied to the adoption of new medical technologies.<sup>1</sup> In this issue of *The American Journal of Managed Care*, Pawlson et al similarly propose that each new *measure* of quality—and each designated data source for a measure—should be tested, and the results of testing published in peer-reviewed journals.<sup>2</sup> In other words, what is the “performance” of performance measures? With the increasing volume of available performance measures—more than 86 measures have been endorsed by the National Quality Forum (NQF) for ambulatory care alone<sup>3</sup>—and the wide-ranging use of performance measures, tests must not only be completed on measures before implementation, but results of tests must be reported in a standardized and meaningful manner to enable selection of the most appropriate measures and data sources for each implementation program.

Pawlson et al make a strong contribution to this field by expanding on previous work to evaluate the use of administrative data for performance measures and in beginning to identify which data elements may be more accurate and complete. The authors’ presentation of results is a starting place for evolving standards for reporting measure-testing results.

## Evaluating Administrative Data: New Wine in an Old Bottle

Some 10 years ago, the Regenstrief Institute for Health Care held a conference on “Measuring Quality, Outcomes, and Cost of Care Using Large Databases.” In a conference paper, lezzoni wrote: “...widespread quality assessment typically demands a tradeoff—the credibility of clinical data versus the expense and feasibility of data collection. Can administrative data produce useful judgments about the quality of healthcare?”<sup>4</sup> Previous studies have questioned the adequacy of administrative data for health services research.<sup>4,5</sup> Pawlson et al take these results a constructive step further by quantifying the differences in performance results based on data source across 15 performance measures using recently collected data.

The authors compare Healthcare Effectiveness Data and Information Set (HEDIS®) performance measure results derived from administrative data alone with administrative data combined with

selective manual chart review, known as hybrid data collection. The National Committee for Quality Assurance (NCQA) specifies a measure for hybrid data collection if preliminary field test performance rates based on administrative data alone vary by more than 5% from the rates achieved with hybrid data collection. For all measures specified for hybrid data collection, health plans may choose to report on those measures using administrative data alone or using the hybrid method. Because health plans are likely to choose the hybrid method when their results using administrative data alone are not adequate, we would expect to see differences above 5%. If the key question of interest concerns the differential quality of data sources, an acid test might be to compare both administrative-only and hybrid data to a gold standard for measures that have not yet undergone field testing.

Noteworthy among the results is the degree and scope of the differences. Performance rates change by more than 10% on 12 of the 15 measures studied during 2 separate years, including 4 widely used measures for diabetes care. Further, the difference in performance rates between the administrative data only and hybrid method varied considerably across health plans. These findings should encourage further research on at least 2 fronts. One natural question is: What accounts for the differences in measurements obtained through administrative data and hybrid data? Pawlson et al rely upon administrative data for denominators in both administrative data only and hybrid data collection methods; thus, it follows that observed differences by data source must be due to differences in numerator and/or exclusion case finding. Second, as the authors note, a concerning implication of the findings is that performance results based on administrative data alone may be even more unstable at the individual provider level. Additional research will be needed to disentangle the effects that data source, small sample sizes, small physician effects, and clustering have on the reliability of individual-level practice profiles.<sup>6,7</sup>

## Reliable Data Elements: Finding the Best of the Breed

Although Pawlson et al do not provide a rigorous decomposition of reliability for specific data elements by data source, their article draws important attention to the heterogeneity in sources for data elements that go into the calculation of performance measures. Administrative data, medical charts,

[www.ajmc.com](http://www.ajmc.com)  
Full text and PDF

and hybrid data may provide different data collection advantages and costs by data element. For example, documentation of performance of screening tests should be reliably captured by claims from laboratories, even if the clinical laboratory results may be less reliably captured across different laboratories. Efforts should be undertaken to optimally match data sources to measures so that the greatest accuracy can be realized at the least data collection costs.

Another approach to improving the reliability and validity of administrative data for quality measurement is to expand and enhance standard coding conventions to capture physician actions and decision making that is indicative of quality. This strategy has been undertaken by the Current Procedural Terminology (CPT®) Editorial Panel, which has developed CPT Category II codes as a means to report more detailed claims data required for performance measure reporting than is possible through CPT codes (Category I or Category III) used for payment of services.<sup>8</sup> The success of this approach will, of course, need to be tested.

#### **Electronic Health Record Systems (EHRS)**

EHRS hold great promise to provide detailed clinical data, but the lack of clinical data standardization across settings and across products remains a challenge for automated data extraction. As a result, the same problems with reliability and validity that limit administrative data-based quality measurement, currently also limit EHRS-based quality measurement.<sup>9-11</sup>

It is unlikely that a single data source will ever be able to satisfy the requirements of low data collection costs, timeliness, universal availability, and inclusion of clinical detail that measure developers and users would like, to say nothing of the desire of physicians and other healthcare professionals to measure results in a timely and usable manner. However, ideal performance measures can be engineered if the data source is considered early in the measure development process.

#### **Standards for Testing and Reporting**

Other genres of medical literature, such as clinical trials,<sup>12</sup> cost-effectiveness analyses,<sup>13</sup> and diagnostic accuracy studies,<sup>14</sup> have evolved standards for reporting in recognition of the role that their literature plays in major decision-making within the healthcare sector. The best data and the best tests cannot produce useful judgments about the quality of healthcare without transparency in the communication of research design and results.

The development or adoption of similar standards could also benefit the emerging field of performance measurement. The NQF intends to promote standards for testing of performance measures by delineating "time-limited" endorse-

ment, whereby measure developers have 2 years in which to produce and provide evidence of testing. The American Medical Association-convened Physician Consortium for Performance Improvement® (Consortium) will vote at its October 2007 meeting on a protocol to guide and define its measure testing activities. Although these initiatives provide a framework to structure measure testing activities, no one has yet broached the problem of ensuring accurate and appropriate reporting of measure testing results. The results described by Pawlson and colleagues provide a few specific illustrations why the field of quality measurement may benefit from a more uniform approach to reporting measurement studies.

For example, the authors present the means of each performance measure derived from each type of data, but omit measures of variability for the 2 measurement distributions although one may well be interested in whether variability of measurements differ by data source. Another example where a recommendation could be useful concerns reporting of rank order changes that would result under alternative data collection strategies. Inadequate description of the distributions within which observations are ranked can obfuscate the meaning of rank order changes. Reporting the number and percent of observations experiencing a change in quartile rank resulting from measurement under the alternative data collection strategies would be more informative with percentile values, and/or interquartile ranges. A shift from one quartile to an adjacent quartile has quite different implications depending on the width of the interquartile range.

Reporting standards also may be valuable in communicating and presenting statistical comparisons. Pawlson et al present means of quality indicators calculated from 2 alternative data collection strategies as well as the differences in means, but do not present the results of formal statistical tests of significance for the differences. Other authors have adapted sensitivity, specificity, and predictive values for use in evaluating the reliability of performance measures across different data collection strategies<sup>15</sup> or reported  $\kappa$  statistics to summarize reliability of measurements across different data sources.<sup>16</sup> While differences in research protocols may favor one form of statistical comparison over another, general reporting standards could help ensure appropriateness and sufficient rigor to inform sound decision-making about the reliability of performance measures.

Administrative data will continue to be the "go-to" source for many implementers given the desire for readily available data from a large population of providers with minimal data collection burdens. Given the concerns raised by Pawlson et al and other researchers, where do we go from here?

First, all stakeholders must understand and acknowledge the limitations of current quality measurement involving

## The Performance of Performance Measures

administrative data; some measures based strictly on administrative data should not be used in particular programs. Second, more testing of quality indicators such as that contributed by Pawlson et al is necessary to better address the usefulness of performance measurement utilizing administrative data as well as other data sources. An effort to standardize testing and reports of test results would also be beneficial. Third, technology must also be leveraged to enhance quality measurement. EHRS hold great promise for performance measurement by virtue of expanding the amount of clinical data available, facilitating reporting, and providing feedback mechanisms for quality measurement. Measure developers have begun to work more closely with EHRS vendors, and the NQF is seeking to assess the reliability of data elements within EHRS for performance measures. These efforts are complementary and good places to start.

**Authors' Affiliation:** From the American Medical Association, Chicago, IL (KSK, JC, SS).

**Author Disclosure:** KSK reports being the principal investigator for a grant funded by the Agency for Healthcare Research and Quality (Greg Pawlson, MD, is among the key personnel for that grant). KSK and SS report collaborating with the NCQA on a measurement development grant funded by the Centers for Medicare & Medicaid Services (SAS is the AMA project manager; KSK is the co-author of the proposal).

**Authorship Information:** Concept and design; drafting of the manuscript, and critical revision of the manuscript for important intellectual content (KSK, JC, SS).

**Address correspondence to:** Karen S. Kmetik, PhD, American Medical Association, 515 N State St, Chicago, IL 60611. E-mail: karen.kmetik@ama-assn.org

---

## REFERENCES

1. Auerbach AD, Landefeld CS, Shojania KG. The tension between needing to improve care and knowing how to do it. *N Engl J Med*. 2007;257:608-613.
2. Pawlson LG, Scholle SH, Powers A. Comparison of administrative-only versus administrative plus chart review data for reporting HEDIS® hybrid measures. *Am J Manag Care*. 2007;13:553-558.
3. National Quality Forum. Standardizing Ambulatory Care Performance Measures. Available at: <http://www.qualityforum.org/projects/ongoing/ambulatory/index.asp>. Accessed September 17, 2007.
4. Iezzoni LI. Assessing quality using administrative data. *Ann Intern Med*. 1997;127:666-674.
5. Maclean JR, Fick DM, Hoffman WK, King CT, Lough ER, Waller JL. Comparison of 2 systems for clinical practice profiling in diabetic care: medical records versus claims and administrative data. *Am J Manag Care*. 2002;8:175-179.
6. Hofer TP, Hayward RA, Greenfield S, et al. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA*. 1999;281:2098-2105.
7. Krein SL, Hofer TP, Kerr EA, Hayward RA. Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. *Health Serv Res*. 2002;37:1159-1180.
8. Current Procedural Terminology® Editorial Panel. *CPT Process—How a Code Becomes a Code*. Available at: <http://www.ama-assn.org/ama/pub/category/3882.html>. Accessed September 17, 2007.
9. Tang PC, Ralston M, Arrigotti MF, Qureshi L, Graham J. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J Am Med Inform Assoc*. 2007;14:10-15.
10. Baker DW, Persell SD, Thompson JA, et al. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Ann Intern Med*. 2007;146:270-277.
11. Persell SD, Wright JM, Thompson JA, Kmetik KS, Baker DW. Assessing the validity of national quality measures for coronary artery disease using an electronic health record. *Arch Intern Med*. 2006;166:2272-2277.
12. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001;134:663-694.
13. Gold MR, Siegel JE, Russell LB, et al. *Cost-Effectiveness in Health and Medicine*. New York, NY: Oxford University Press; 1996.
14. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003;326:41-44.
15. Benin AL, Vitkauskas G, Thornquist E, et al. Validity of using an electronic medical record for assessing quality of care in an outpatient setting. *Med Care*. 2005;43:691-698.
16. Kerr EA, Smith DM, Hogan MM, et al. Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures. *Jt Comm J Qual Improv*. 2002;28:555-565. ■