EDITORIAL

Evaluating Quality in Small-Volume Hospitals

An unsolved question in health care quality measurement is how to assess the performance of hospitals and physicians with small volumes of patients.[1] Evaluating hospitals with limited information is akin to judging a baseball player's hitting performance based on only a few games. Major League Baseball publishes all players' batting averages,[2] but it considers only players with at least 502 plate appearances over a season when ranking top individual batting performances.[3] Unlike most baseball players, many hospitals with small volumes may never accumulate enough data to support an accurate evaluation. What should we do?

CHALLENGES OF ASSESSING HOSPITAL PERFORMANCE

In this issue of the *Archives*, O'Brien et al[4] illustrate the challenges of assessing hospital performance when case volumes are small. They examine the relationship between hospital performance and sample size using the process-of-care measures of quality for acute myocardial infarction (AMI) published by the Centers for Medicare and Medicaid Services (CMS) and the Hospital Quality Alliance.[5] These measures estimate quality by the proportion of eligible patients who received the recommended treatment—for example, the percentage of patients with AMI who received aspirin on arrival at the hospital. Measure performance is published on the Hospital Compare Web site,[6] where each hospital's results are presented as percentages displayed in bar graphs. The Web site uses benchmarking to define "top hospital" performance—that is, each hospital's result is compared with those of its best-performing peers (specifically, the 90th percentile of performance for each individual measure). Because many hospitals treated all eligible patients who were sampled, the top hospital benchmark was set to a perfect score for most measures.

See also page 1277

This approach seems straightforward enough, but O'Brien et al[4] demonstrate that it also makes it relatively difficult for high-performing large hospitals to become top performers and relatively easy for low-volume hospitals to achieve the top hospital designation. They also show how using an alternative statistical model, the hierarchical generalized linear model (HGLM), can have the opposite effect. The implications of both statistical approaches on small hospitals merit discussion because the CMS mortality measures reported on the Hospital Compare Web site are based on the HGLM, unlike the process measures.

The problem with the top hospital designation is perhaps easiest to understand using another baseball analogy. In the 2007 regular season of the Major League, 10 players had perfect batting averages of 1.000.[2] Yet the lifetime batting averages of these players (predominantly pitchers) were not substantially higher than those of their teammates. The reason for this finding, of course, is that all of the players batting 1.000 had 3 or fewer at-bats; it is the small sample size that makes it possible to achieve "extraordinary performance."

A similar phenomenon occurs with hospital benchmarking. Consider a group of hospitals, all with the same long-term prescription rate for aspirin on arrival for AMI as the national average (95% of eligible patients). A hospital with 10 cases in a reporting period has an approximately 60% chance of perfect performance and recognition as a top hospital; a larger hospital with 100 cases has less than a 1% chance of perfect performance. Now consider a much poorer performing hospital that prescribes aspirin for just 50% of patients but has a very small volume of 5 cases; such a hospital is more likely to be assigned top hospital status than the larger hospital with better underlying performance. Estimates associated with small hospitals are "noisy."

The consequence of this approach is that hospitals with average or even poor long-term performances stand to have a higher chance of achieving the benchmark if they have relatively few patients compared with larger volume medical centers.[7] Hospital Compare[6] attempts to mitigate this problem by excluding hospitals with fewer than 25 eligible cases from the bar graphs and from the top hospital designation. But as O'Brien et al[4] demonstrate, this phenomenon can persist even among hospitals with more than 25 cases.

Just as small-volume hospitals are disproportionately represented as top hospitals, the opposite could also occur. Using our baseball analogy again, in the 2007 season, 172 of 971 players in the Major League had batting averages of zero; most of these players had fewer than 10 at-bats over the season.[2] In a similar way, many small-volume medical centers are likely to be found in the lowest percentiles of performance.

The take-home message from O'Brien et al[4] is not that small hospitals are unfairly crowned top performers—rather, it is that with limited information, small hospitals are more at risk of being misclassified at either extreme of performance. When estimates are noisy, the public may be misled toward believing that quality measures are more reliable than they truly are. For this reason, hospital volumes or an estimate of the uncertainty in the measures should be provided.

O'Brien et al[4] then explore how hospital rankings would change after applying alternative statistical approaches to the same data. The HGLM technique deals with a number of important concerns associated with profiling, including accounting for correlation among patient observations treated in the same hospital, providing hospital-specific performance estimates, and handling noisy estimates. It is the statistically appropriate method for measuring hospital mortality rates given the hierarchical nature of the data and the need to risk adjust for differences in hospital case mix.[8,9]

Importantly, the implications for small hospitals are different under HGLM compared with proportions. The HGLM model incorporates a Bayesian concept that, in the absence of information, the most reasonable estimate of a particular hospital's performance is the mean (pretest probability); as more data become available that demonstrate better or worse performance (ie, with more eligible patients) the resulting estimate shifts away from the average (posttest probability).[8] As a result, estimates for hospitals with small samples are more likely closer to the mean. This approach provides a better estimate of their real rate. But, as O'Brien et al[4] illustrate, the HGLM model makes it harder for hospitals with small samples to distinguish themselves as top hospitals. It is just as important to note that HGLM makes it more difficult for small medical centers to be inappropriately labeled as poor performers.

These examples demonstrate that performance measurement reports for small-volume hospitals can be uninformative or, worse, even misleading. We could, therefore, choose to simply not report them. But these hospitals comprise a considerable proportion of the hospitals examined on the Hospital Compare Web site[6]—one-quarter to one-half of hospitals report fewer than 25 cases in the process measures for AMI. Small-town hospitals have had a long tradition of providing for their community's medical care. They should not be shut out of performance rewards, nor should their efforts to pursue quality improvement be discouraged. And occasionally small numbers do contain valuable information; for example, a routine procedure incurring 3 deaths in a row may not necessarily be statistically significant but may suggest the need for a review.

HOW SHOULD WE PROCEED?

Increasing the number of cases by raising minimum sampling requirements or lengthening the sampling period is 1 solution. The recent report of the CMS to Congress on value-based purchasing recommends steps in this direction, such as increasing sample size requirements.[10] However, collecting additional data will cost hospitals money. Moreover, longer sampling periods mean that reported performance lags behind actual performance. Hospitals may find it difficult to demonstrate improvement if weighed down by suboptimal quality in prior years, and recent quality declines may be masked by prior excellence. Still, these trade-offs are worth considering.

Another solution is to use composite measures. Well-designed composite measures can provide incentives to smaller medical centers to improve quality while relieving pressure to game results by declining high-risk patients. But composite measures also can have unintended consequences. A recent evaluation of a CMS "pay for performance" demonstration project for coronary artery bypass surgery found that it was possible for a hospital with an extremely low death rate to score poorly on the composite measure, potentially triggering inappropriate financial penalties.[11] Although the problem can be addressed by weighting and standardizing the subcomponents, deciding what weights to assign is complex.[12]

Regardless, we should report the results we generate for small-volume hospitals more carefully. Most important, we should convey the uncertainty associated with the results, preferably in a way that consumers understand. That is hard to do, but research on consumers' use of health information provides some guidance. For example, it shows that consumers are better able to interpret frequencies than percentages.[13,14] Baseball cards are consistent with this approach; they present both the numerator (hits) and denominator (at bats) alongside the percentage (batting averages). For process measures, CMS should consider presenting the number of patients at each hospital with the percentage who receive the desired treatment. These are currently available on a drill-down menu, but one has to look for them. Reporting the probability that a hospital is a top performer is also possible; interestingly, HGLM was popularized by an analysis to determine the probability that Ty Cobb was really a .400 hitter.[15]

We could provide other quantitative estimates of uncertainty; for example, CMS calculates 95% interval estimates (like confidence intervals) for mortality measures. But interval estimates are relatively difficult for consumers to interpret. For public reporting, the CMS translates the rates and interval estimates into 3 categories on the Hospital Compare Web site[6] by classifying hospitals as "better than the U.S. national rate," "worse than the U.S. national rate," or "no different than the U.S. national rate." Most small-volume hospitals end up in the latter category because of their size, but consumers may incorrectly conclude that a hospital that is "no different" is an average performer. Again, reporting the volume of cases would provide much-needed context for interpreting the results.

Qualitative descriptions of high or low performance that assist consumers are problematic but necessary. They inevitably reflect policy choices (90th percentile of performance for process measures or 95th percentile interval estimate for mortality measures) and can hide small-volume effects. But we should not abandon them; research shows that if hospital quality information is too complex, consumers are less likely to use it in their decisions.[14] Rather, we should state the rules for such labels clearly and better anticipate unintuitive effects.

Ultimately, no easy or simple solution exists for the reporting of small-sample hospitals. Policymakers are faced with multiple tradeoffs when publicly reporting measures with small samples: accuracy, complexity, accessibility, reliability, fairness, and timeliness. As more payers move toward pay-for-performance and as performance measurement extends from hospitals to individual physicians, policymakers will continue to look for strategies to distinguish among health care providers serving relatively small numbers of patients. These efforts will serve quality improvement as long as the reported results accurately reflect what the data can and cannot tell us.

Elizabeth E. Drye, MD, SM
Jersey Chen, MD, MPH

Correspondence: Dr Drye, Yale/Yale–New Haven Hospital Center for Outcomes Research and Evaluation, 1 Church St, Ste 200, New Haven, CT 06510-3330 (elizabeth.drye@yale.edu).

Financial Disclosure: Drs Drye and Chen are currently subcontractors for the CMS through a contract with the Colorado Foundation for Medical Care regarding the design and implementation of publicly reported quality measures.
Additional Contributions: Sharon-Lise T. Normand, PhD, and Harlan M. Krumholz, MD, SM, provided helpful comments, and Kanchana Bhat, MPH, provided assistance with the preparation of this manuscript.

Disclaimer: The views expressed in this editorial are the authors' own.

REFERENCES

1. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. *JAMA*. 2004;292(7):847-851.
2. Major League Baseball Web site. http://mlb.mlb.com/stats/historical/player_stats.jsp. Accessed January 8, 2008.
3. 2007 Official baseball rules. Rule 10.22. Major League Baseball Web site. http://www.mlb.com. Accessed January 8, 2008.
4. O'Brien SM, DeLong ER, Peterson ED. Impact of case volume on hospital performance assessment. *Arch Intern Med*. 2008;168(12):1277-1284.
5. Jha AK, Li Z, Orav EJ, Epstein AM. Care in U.S. hospitals: the Hospital Quality Alliance program. *N Engl J Med*. 2005;353(3):265-274.
6. Hospital Compare Web site. http://www.hospitalcompare.hhs.gov. Accessed January 8, 2008.
7. Dimick JB, Welch HG. The zero mortality paradox in surgery. *J Am Coll Surg*. 2008;206(1):13-16.
8. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med*. 1997;127(8, pt 2):764-768.
9. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg*. 2001;72(6):2155-2168.
10. Report to Congress: Plan to Implement a Medicare Hospital Value-Based Purchasing Program, 2007. US Department of Health and Human Services, Centers for Medicare and Medicaid Services, Web site. http://www.cms.hhs.gov/AcuteInpatientPPS/downloads/HospitalVBPPlanRTCFINALSUBMITTED2007.pdf. Accessed January 8, 2008.
11. O'Brien SM, DeLong ER, Dokholyan RS, Edwards FH, Peterson ED. Exploring the behavior of hospital composite performance measures: an example from coronary artery bypass surgery. *Circulation*. 2007;116(25):2969-2975.
12. Tu JV, Austin PC. Cardiac report cards: how can they be made better? *Circulation*. 2007;116(25):2897-2899.
13. Hibbard JH, Peters E. Supporting informed consumer health care decisions: data presentation approaches that facilitate the use of information in choice [published online November 6, 2001]. *Annu Rev Public Health*. 2003;24:413-433.
14. Peters E, Dieckmann N, Dixon A, et al. Less is more in presenting quality information to consumers. *Med Care Res Rev*. 2007;64(2):169-190.
15. Morris CN. Parametric empirical Bayes inference: theory and applications. *J Am Stat Assoc*. 1983;78(381):47-55.