Impact of Case Volume on Hospital Performance Assessment

Sean M. O'Brien, PhD; Elizabeth R. DeLong, PhD; Eric D. Peterson, MD, MPH

Background: Process performance measures are increasingly used to assess and reward hospital quality. The impact of small hospital case volumes on such measures is not clear.

Methods: Using data from the Hospital Quality Alliance, we examined hospital performance for 8 publicly reported process measures for acute myocardial infarction (AMI) from 3761 US hospitals during the reporting period of January to December 2005. For each performance measure, we examined the association between hospital case volume, process performance, and designation as a "top hospital" (performance at or above the 90% percentile score).

Results: Sample sizes available for process performance assessment varied considerably, ranging from a median of 3 patients per hospital for timely administration of thrombolytics therapy to 62 patients for aspirin given on arrival at the hospital. In aggregate, hospitals with larger AMI case volumes had better process performance; for example, use of β-blockers at arrival rose from 72% of patients at hospitals with less than 10 AMI cases to 80% of patients at hospitals with more than 100 cases ($P < .001$ for volume trend). In contrast, owing to an artifact of wide sampling variation in sites with small denominators, classification of a center as a top hospital actually declined rapidly with increasing case volume using current analytic methods ($P < .001$). This unexpected association persisted after excluding very low volume centers (<25 cases) and when using Achievable Benchmarks of Care. Using hierarchical models removed the paradoxical association but may have introduced a bias in the opposite direction.

Conclusions: Large-volume hospitals had better aggregate performance but were less likely to be identified as top hospitals. Methods that account for small and unequal denominators are needed when assessing hospital process measure performance.

Arch Intern Med. 2008;168(12):1277-1284

Over the past decade, performance measurement has become an integral part of the strategy to narrow the quality chasm and improve the nation's health care. Quality measures are increasingly used by patients and payers to evaluate and compare hospitals. To encourage excellent performance, producers of reports often benchmark hospital results relative to "top medical centers," defined as the 10% with the highest performance achieved. Hospitals designated as best-performing medical centers gain public recognition and often stand to receive direct financial rewards under pay-for-performance programs.[1-3]

For editorial comment see page 1249

It is common knowledge that performance measures are subject to chance fluctuations and are more susceptible to these fluctuations when the sample sizes are small. However, the magnitude and impact of these chance fluctuations may be underestimated by users of performance measures. We speculated that existing simple ranking methods that are used in public reporting would lead to spurious inferences when identifying top hospitals. Furthermore, because the precision of a performance measure is proportional to the number of patients assessed, we conjectured that small and large hospitals would not be compared on a level playing field.

An important case study for illustrating these issues and their policy implications can be found in the US Centers for Medicare & Medicaid Services (CMS) and the Hospital Quality Alliance (HQA) Improving Care Through Information (ICTI) initiative. The HQA-ICTI initiative is a large-scale, public and private hospital collaboration that seeks to make performance information available for all acute-care nonfederal hospitals. Participating

Author Affiliations: Duke Clinical Research Institute (Drs O'Brien, DeLong, and Peterson), Department of Biostatistics and Bioinformation (Drs O'Brien and DeLong), and Division of Cardiology (Dr Peterson), Duke University Medical Center, Durham, North Carolina.

Table 1. Distribution of Hospital Sample Sizes and Overall Performance for AMI Performance Measures

Variable	ACE or ARB Prescribed for LSVD[a]	Aspirin Given at Arrival	Aspirin Prescribed at Discharge	β-Blocker Given at Arrival	β-Blocker Prescribed at Discharge	PCI Administered Within 120 Min[b]	Smoking Cessation Counseling	Thrombolytics Given in ≤30 min
Hospitals, No.								
1-9	1443	645	1034	687	1001	158	1225	1388
10-24	605	561	619	596	603	290	344	261
25-39	320	305	328	337	337	340	199	55
40-99	572	827	506	926	541	472	512	11
≥100	266	1404	1156	1178	1167	39	478	0
Cases, No.								
Overall[c]	3206	3742	3643	3724	3649	1299	2758	1715
Per hospital, median[d]	13	62	31	53	33	33	14	3
MSD	20	75	80	43	66	9	19	7
Hospitals not meeting MSD, %	60	54	65	46	61	12	66	46
National usage rate, %[e]	83	95	96	92	95	69	92	38

Abbreviations: ACE, angiotensin-converting enzyme; AMI, acute myocardial infarction; ARB, angiotensin II receptor blocker; LSVD, left systolic ventricular dysfunction; MSD, minimum sufficient denominator; PCI, percutaneous coronary intervention.

[a] Denotes use of ACE inhibitors or ARBs for eligible patients with LSVD.

[b] Denotes percutaneous coronary intervention received within 120 minutes of arrival at hospital.

[c] Number of hospitals with at least 1 eligible case.

[d] Excluding hospitals with no eligible cases.

[e] The fraction of patients who received the care process among all eligible patients in the entire database.

hospitals report process-of-care measures pertaining to acute myocardial infarction (AMI), heart failure, community acquired pneumonia, and the prevention of surgical infections. Hospitals scoring at or above the 90th percentile for a measure are described as top hospitals on their publicly available Web site.[4]

Focusing on 8 core hospital quality indicators for AMI, we first examined the association between hospital sample sizes and observed performance on individual process-of-care quality measures. We then analyzed the association between hospital volume and top hospital designation using a probabilistic model and actual empirical results from the HQA-ICTI initiative. Third, we examined the degree to which these associations changed if one used alternative analytic strategies including (1) deleting sites with fewer than 25 cases, (2) using the Achievable Benchmarks of Care (ABC)[5] methods to determine benchmarks, or (3) using hierarchical models to estimate hospital performance.

METHODS

DATA SOURCE AND PERFORMANCE MEASURES

Details of the HQA-ICTI public reporting initiative and measure specifications are described elsewhere.[6] Performance data were downloaded from the Hospital Compare Web site[4] for the reporting period January to December 2005.We initially examined the association between case volume and process performance and performance ratings for all 20 performance measures that were reported on the Hospital Compare Web site[4] during 2005. Given the substantial similarity in findings among these results, we ultimately used the 8 process-of-care measures for AMI to illustrate these findings. The names of these measures can be found in **Table 1** and **Table 2**. Specific definitions and inclusion and exclusion criteria for these individual AMI measures can be found at the Hospital Compare Web site.[4]

A hospital's score for a performance measure is defined as the percentage of eligible patients who received the indicated care process rounded to the nearest whole number. Hospitals with scores that meet or exceed the national 90th percentile are described as "top hospitals" on the Hospital Compare Web site.[4]

STATISTICAL ANALYSIS

To assess the statistical precision of the selected performance measures, we first examined the distribution of hospital-specific denominators across the 8 measures and 3761 hospitals. For comparison, we also calculated the minimum sufficient denominator (MSD) for each measure using a formula suggested by Kiefe et al.[5] The MSD for a given process measure is defined as the smallest number *N* such that 100% adherence attained on a denominator of *N* patients would be statistically different from the overall mean performance on the process measure. Kiefe et al[5] state that when the denominator for a center falls below the MSD, including that center in the calculation of a benchmark runs the risk of distorting the overall performance and inflating the benchmark.[5]

RELATIONSHIP BETWEEN SAMPLE SIZE AND HOSPITAL PROCESS PERFORMANCE AND TOP HOSPITAL RATING

To illustrate the association between volume and performance, hospitals submitting at least 1 case were grouped into 5 patient volume categories based on the number of eligible cases. The categories were 1-9, 10-24, 25-39, 40-99, and 100 or more eligible patients. For each measure, variation across the categories was assessed by calculating the proportion of patients within each volume category who received the recommended care process. This proportion was calculated by aggregating data across hospitals to form a single aggregate numerator and denominator. To study the association between sample size and attainment of the top hospital designation, we calculated the proportion of hospitals in each

Table 2. Performance Benchmarks for AMI Performance Measures

Benchmark	ACE or ARB Prescribed for LSVD	Aspirin Given at Arrival	Aspirin Prescribed at Discharge	β-Blocker Given at Arrival	β-Blocker Prescribed at Discharge	PCI Administered Within 120 Min	Smoking Cessation Counseling	Thrombolytics Given in ≤30 min
90th Percentile of raw performance scores[a]	100	100	100	100	100	88.0	100	76.0
90th Percentile of raw performance scores[b]	96.0	100	99.0	99.0	100	87.0	100	71.0
ABC benchmarks, %[c]	97.7	99.7	99.9	99.4	99.6	91.6	100	80.1
90th Percentile of the hierarchical estimates	90.6	98.2	98.2	97.3	97.9	82.7	98.1	47.5

Abbreviations: ABC, Achievable Benchmarks of Care methodology (Kiefe et al[5]); ACE, angiotensin-converting enzyme; AMI, acute myocardial infarction; ARB, angiotensin II receptor blocker; LSVD, left systolic ventricular dysfunction; PCI, percutaneous coronary intervention.

[a] Hospitals meeting or exceeding this threshold are top hospitals according to the Hospital Compare Web site.[4]

[b] Calculated among hospitals with at least 25 patients who were eligible to receive the indicated care.

[c] Using the method of Kiefe et al[5,7] and Weissman et al.[9]

volume category that met or exceeded the reported national 90th percentile (ie, were top hospitals).

PROBABILITY CALCULATIONS

In addition to analyzing the empirical (observed) relationship between volume and top hospital status, we also performed probability calculations to determine the effect of small sample size on a hospital's likelihood of being designated a top hospital. Based on data from the HQA-ICTI initiative, we observed that the threshold performance for top hospital designation equaled 100% adherence for 6 of the 8 process measures (all except percutaneous coronary intervention [PCI] administered within 120 minutes and thrombolytics given within 30 minutes). We thus considered the probability that a hospital would attain 100% adherence on a given process measure. This probability was calculated according to the binomial distribution with the following formula:

$$\text{Probability That 100\% of Patients Receive Care Process} = P^N,$$

where P denotes the probability that the care process is delivered to any single patient and N denotes the number of patients. This probability formula was evaluated for various combinations of N and P to determine the importance of sample size on the probability of being classified as a top hospital. Because P is less than 1.0, it is clear that, even if every hospital and every patient have the same P, those with higher case loads have a smaller probability of achieving a perfect performance on the measure.

PERFORMANCE OF ALTERNATIVE BENCHMARK METHODS

We also explored 3 alternative methods of determining benchmarks and identifying top hospitals. First, we simply excluded hospitals having fewer than 25 eligible patients when calculating top hospital benchmark performance rates. We chose this threshold because the Hospital Compare Web site warns users that sample sizes of less than 25 are "too small for purposes of reliably predicting hospital's performance."[4] Despite this warning, all hospitals, regardless of sample size, are included in the calculation of the reported national 90th percentile. Second, we determined benchmarks by applying the ABC method proposed by Kiefe et al[5,7] and Weissman et al.[9] This method uses a Bayesian estimator to reduce the influence of small sample sizes and produces a benchmark that is interpreted as "the mean of the best care achieved for at least 10% of the population."[5(p444)] The ABC method has the advantage of reducing the influence of small denominators while using all of the available data in the benchmark calculation rather than simply eliminating hospitals with few eligible patients.

Finally, we used a hierarchical model to calculate the empirical Bayes estimate of each hospital's true (long-run) usage rate for each measure. Hierarchical models combine information from all hospitals when estimating the usage rate for a single hospital, thereby borrowing strength from the ensemble to obtain a more stable estimate.[10,11] This approach is advocated for producing estimates that are more reliable reflections of a hospital's likely true performance rate over time and is now becoming the standard for analysis of patient data that form clusters within hospitals.[10,11] Hierarchical modeling is currently used to analyze the new mortality measures that are publicly reported on the Hospital Compare Web site.[4] Each hospital's estimate is "shrunken" toward the overall mean of all hospitals, with the amount of shrinkage being greater for hospitals with smaller sample sizes. This shrinkage property makes the estimates more stable and prevents wide unrealistic fluctuations in sites with small sample sizes. The national 90th percentile was calculated from the distribution of these shrunken hierarchical estimates. The top hospital classification was then based on the shrunken hierarchical estimates rather than on the raw unadjusted data.

RESULTS

DISTRIBUTION OF HOSPITAL-SPECIFIC SAMPLE SIZES

The number of hospitals reporting AMI measures ranged from 1299 for PCI administered within 120 minutes to 3742 for aspirin given at arrival (Table 1). The median number of cases per hospital eligible for individual process measures ranged from 3 patients per hospital for thrombolytics given within 30 minutes to 62 patients per hospital for aspirin given on arrival. For each measure, at least 32% of hospitals had fewer than 25 eligible patients. Percutaneous coronary intervention administered within 120 minutes was the only measure to have a large percentage of hospitals (88%) with sample sizes exceeding the MSD recommended by Kiefe et al[5,7] and Weissman et al.[9] For other measures, at least 46% of hos-

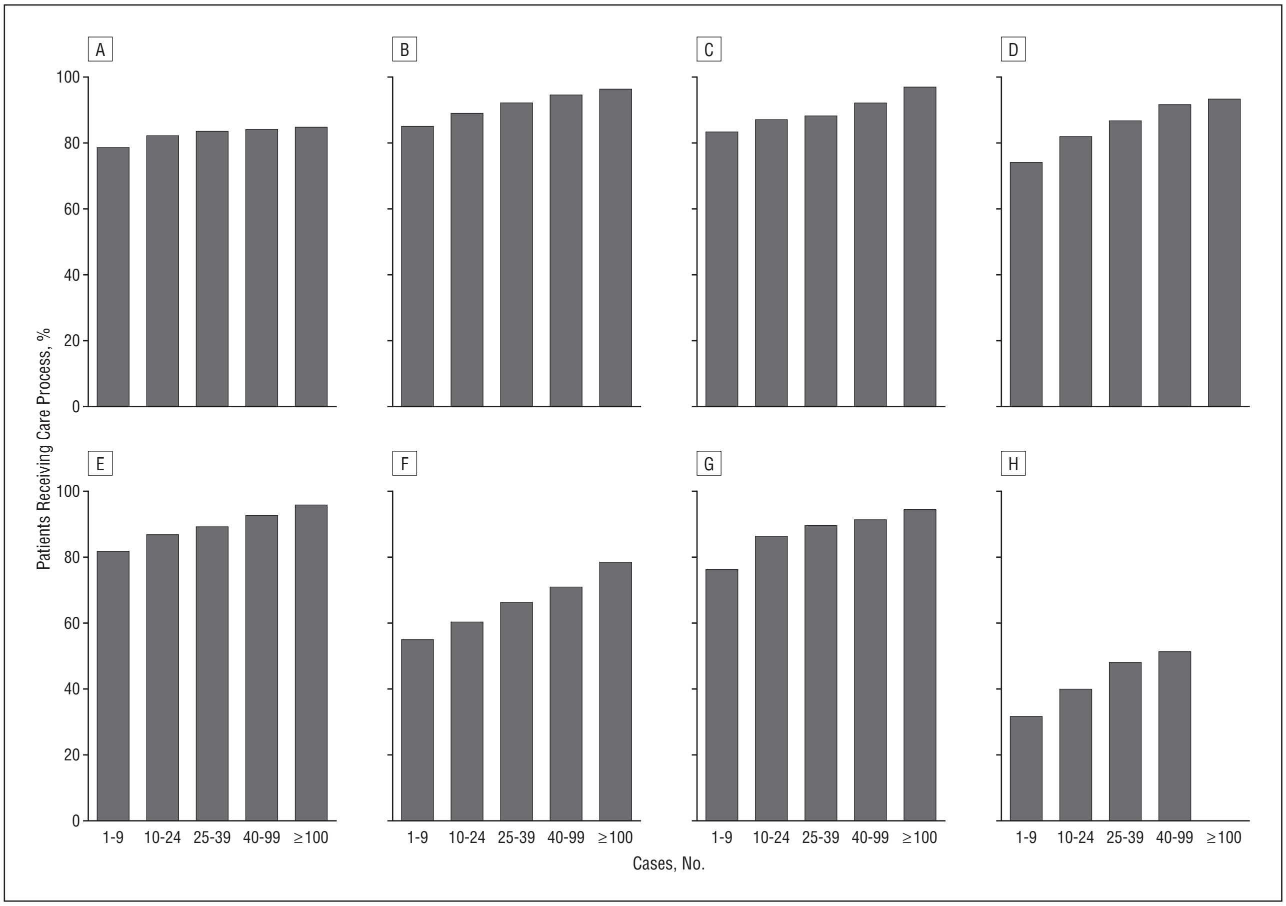

Figure 1. Percentage of patients receiving evidence-based care processes by hospital volume category. A, Angiotensin-converting enzyme or angiotensin II receptor blocker given for left systolic ventricular dysfunction; B, aspirin given on arrival at the hospital; C, aspirin prescribed at discharge; D, β-blocker given at arrival; E, β-blocker prescribed at discharge; F, percutaneous coronary intervention administered within 120 minutes; G, smoking cessation counseling; H, thrombolytic administered within 30 minutes.

pitals fell below the recommended MSD threshold. Only the aspirin and β-blocker measures had a large percentage (≥30%) of hospitals with at least 100 eligible cases. The overall national adherence rates ranged from 38% for timely administration of thrombolytics to 96% for aspirin prescribed at discharge and was greater than 90% for 5 of the 8 performance measures (Table 1).

VOLUME-PROCESS RELATIONSHIP

For each measure, there was a monotonically increasing association between hospital volume category and the percentage of patients receiving the recommended therapy (**Figure 1**). The magnitude of the difference in performance between lowest vs highest volume categories varied by measure as follows: angiotensin-converting enzyme (ACE) or angiotensin II receptor blocker (ARB) for left systolic ventricular dysfunction, 74% vs 84%; aspirin given at arrival, 83% vs 96%; aspirin prescribed at discharge, 82% vs 97%; β-blocker given at arrival, 74% vs 94%; β-blocker prescribed at discharge, 78% vs 96%; PCI administered within 120 minutes, 56% vs 75%; smoking cessation counseling provided, 64% vs 95%; and thrombolytics given within 30 minutes, 20% vs 45%.

TOP HOSPITAL BENCHMARKS

For 6 of the 8 measures, the threshold for defining top hospitals was equal to 100% (Table 2). However, the calculation of this threshold included a large number of hospitals with extremely small denominators. Among hospitals having a perfect performance on ACE inhibitors, for example, over 30% of these hospitals attained a perfect performance by successfully treating a single eligible patient. For other measures, at least 46% of hospitals that had a perfect performance had denominators smaller than 10, and fewer than 15% of hospitals with a perfect performance had sample sizes larger than 100. Because of the large number of hospitals with perfect performance and small denominators, the top hospital threshold may overestimate the level of performance that would be maintained consistently by the nation's top hospitals if performance was assessed on a large number of patients.

EFFECT OF SAMPLE SIZE ON ATTAINMENT OF TOP HOSPITAL BENCHMARK

Ignoring the variation in sample sizes can lead to spurious and counterintuitive conclusions when raw performance scores are used to identify top hospitals. In con-

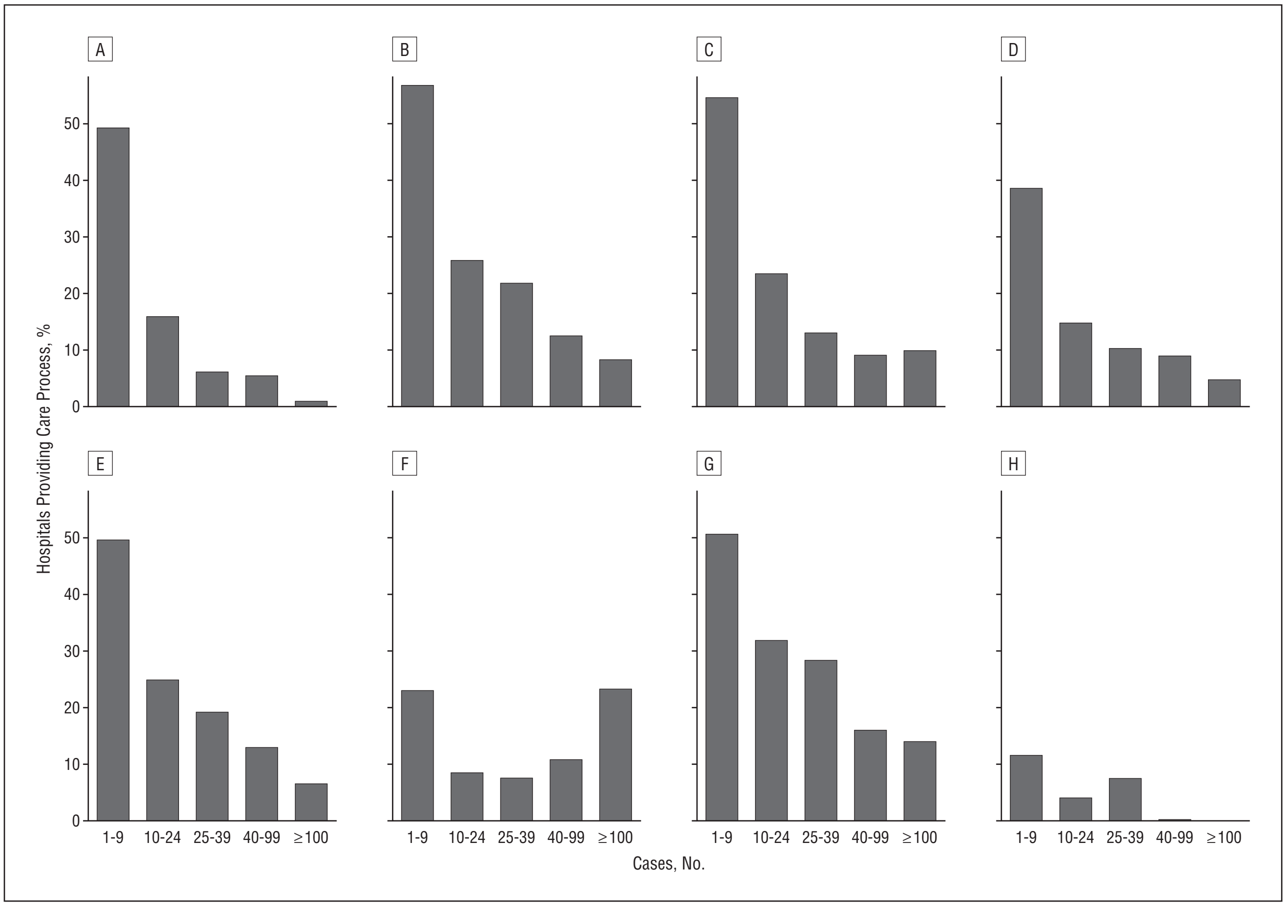

Figure 2. Percentage of hospitals attaining top hospital benchmark by volume category. A, Angiotensin-converting enzyme or angiotensin II receptor blocker given for left systolic ventricular dysfunction; B, aspirin given on arrival at the hospital; C, aspirin prescribed at discharge; D, β-blocker given at arrival; E, β-blocker prescribed at discharge; F, percutaneous coronary intervention administered within 120 minutes; G, smoking cessation counseling; H, thrombolytic administered within 30 minutes.

trast to the directly observed, case volume–process performance relationship illustrated in Figure 1, we found an inverse association between hospital case volume and the percentage of hospitals that attained top hospital status (**Figure 2**). Hospitals with few eligible patients were more likely to achieve a perfect performance because there were fewer opportunities to "fail." The association between volume and top decile performance was not limited to hospitals with fewer than 25 eligible cases. For example, hospitals with 25 to 39 eligible cases were more likely to be top hospitals for 7 of the 8 performance measures vs those centers with at least 40 cases (Figure 2).

PROBABILITY CALCULATIONS

Although chance variation tends to benefit small-volume hospitals when the scores are used to identify top hospitals, it would harm them if the scores were used to identify hospitals with exceptionally poor performance. **Figure 3** presents probability calculations that illustrate the bias against large hospitals when perfect performance is required to be a top hospital. First, irrespective of sample size, hospitals' likelihood of obtaining top hospital status rises as their "true process performance rate" rises, which is reassuring. However, what is less obvious but equally clear is that a hospital's probability of being classified as an exceptional center is also strongly influenced by sample size. If 2 different hospitals have the same true underlying success rate but widely different sample sizes, the smaller hospital is more likely to meet this benchmark. For example, for 2 hospitals with the same true process performance rate of 97%, one with 10 eligible patients and the other with 200 eligible pa-

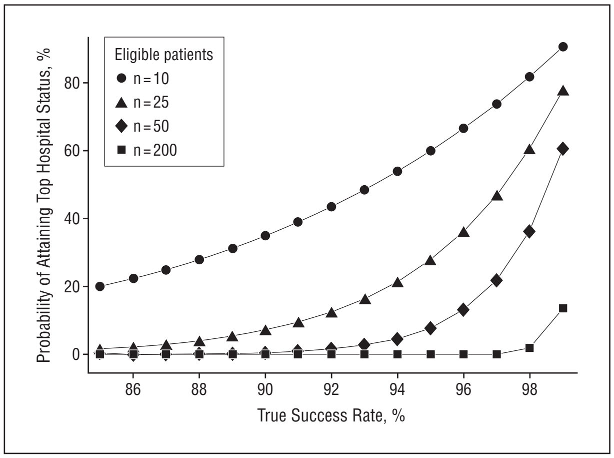

Figure 3. Probability of attaining top hospital benchmark of 100% adherence as a function of sample size and true underlying success probability.

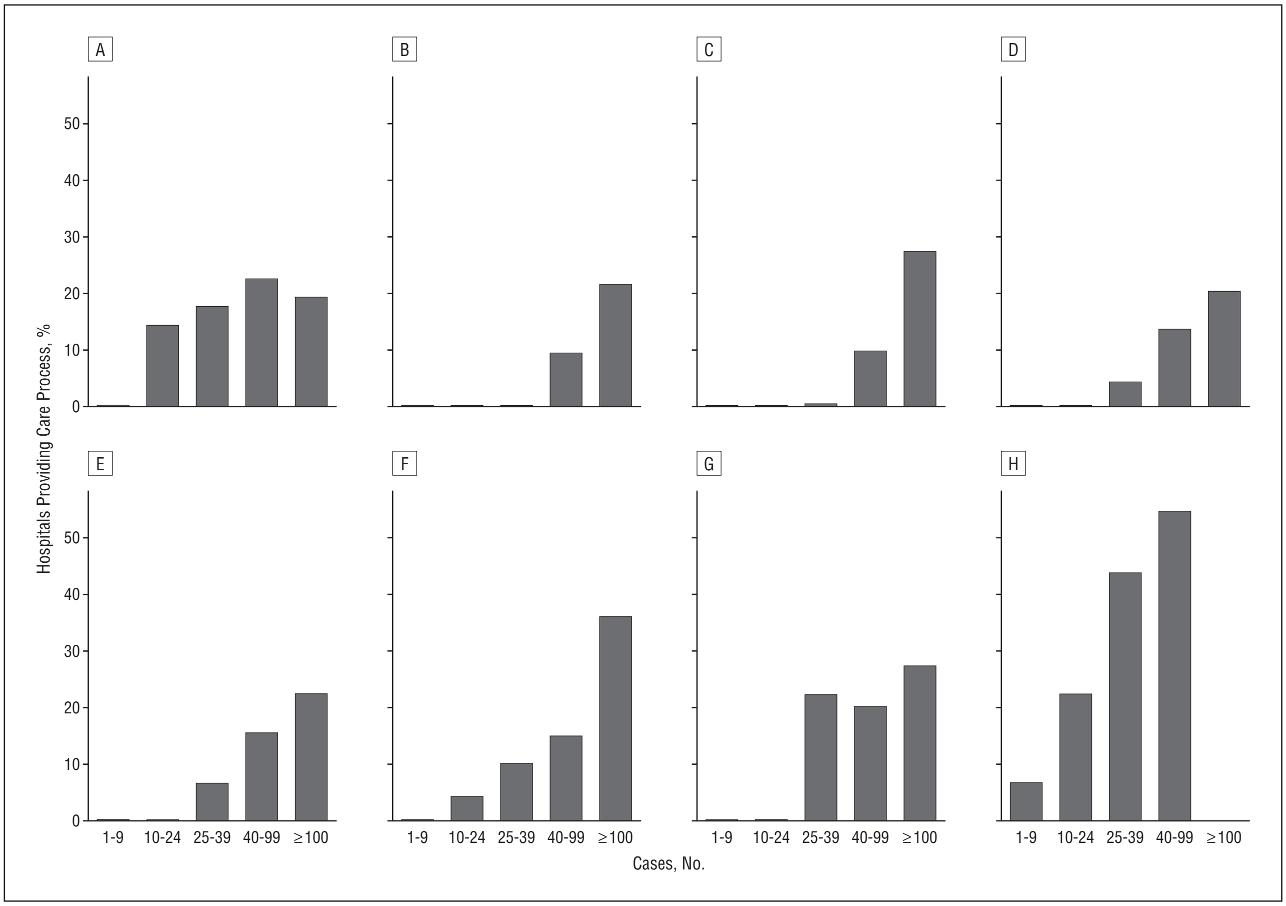

Figure 4. Percentage of hospitals attaining 90th percentile benchmark by volume category when hospital performance is estimated via hierarchical models. A, Angiotensin-converting enzyme or angiotensin II receptor blocker given for left systolic ventricular dysfunction; B, aspirin given on arrival at the hospital; C, aspirin prescribed at discharge; D, β-blocker given at arrival; E, β-blocker prescribed at discharge; F, percutaneous coronary intervention administered within 120 minutes; G, smoking cessation counseling; H, thrombolytic administered within 30 minutes.

tients, the smaller hospital will be over 300 times more likely to achieve top hospital status relative to the larger one. From another perspective, a hospital treating 50 cases and having a true performance rate of 97% would have the same probability of achieving top hospital status as a hospital treating 10 cases with a true performance of only 85%.

ALTERNATIVE BENCHMARK METHODS

One method of dealing with the issue of low case volumes is simply to exclude these from consideration when calculating top hospital ratings. However, after deleting hospitals with fewer than 25 cases, the likelihood of being a top hospital was still highest among the lowest volume hospitals (25-50 cases) for 6 of the 8 measures.

Using the ABC method to determine performance benchmarks also did not change the study's main empirical results. For each measure, the proportion of hospitals meeting the ABC benchmark was greater for hospitals in the lowest category of volume compared with the highest. The calculated benchmark was equal to 100% for 1 measure (smoking cessation) and was greater than 99% for 4 additional measures (aspirin given at arrival; aspirin prescribed at discharge; β-blockers given at arrival; β-blockers prescribed at discharge) (Table 2). The developers of the ABC method do not recommend providing feedback based on ABC benchmarks when the benchmark is greater than 99% because such a measure should be regarded as a standard of care "in which case a less than perfect performance is inappropriate."[12] Thus, the ABC benchmark is arguably not useful for 5 of the 8 AMI process measures.

A third alternative method of dealing with the issue of low case volumes is to estimate performance using hierarchical models. These shrunken hierarchical estimates led to benchmarks that were less than 99% for all 8 process measures (Table 2). In contrast to raw estimates, there was a direct increasing association between hospital volume and the proportion of hospitals that were top hospitals (ie, ranked among the 10% best based on hierarchical estimates) (**Figure 4**). This increasing association between volume and top hospital classification is consistent with the volume-performance association discussed in the "Volume-Process Relationship" subsection (Figure 1).

COMMENT

Process performance assessment has become an integral part of medical practice, yet such assessments should be accurate in order to be meaningful. To date, little attention has been paid to the analytical challenges created by small and unequal sample sizes when measuring hospital process performance. In this study, we found that one-third of US hospitals had fewer than 25 cases for all of their CMS AMI process performance measures. As a result, their performance estimates were highly variable. Although there have been previous calls for standards for publicly reporting risk-adjusted outcome measures,[13,14] our data show that the analysis and reporting of process measures is also challenging.

In this study, we found that the proportion of patients receiving evidence-based care processes at low-volume hospitals was consistently lower than the proportion receiving these processes at high-volume hospitals. This volume-performance trend was observed for each of the 8 AMI process measures and seemed to be monotonic (ie, an increasing trend) across all of the volume categories. This observation is consistent with previous studies[15-17] that have demonstrated a similar direct association between higher hospital volume and lower AMI risk-adjusted mortality rates. Possible mechanisms include more experienced staff at high-volume hospitals (ie, practice makes perfect) and the allocation of resources for continued quality improvement at high-volume hospitals.

Although there seems to be a direct volume-process performance relationship for AMI process measures, the attainment of top hospital status was paradoxically much more common among low-volume sites than high-volume medical centers. This apparent paradox is due to the fact that dichotomous performance measures are subject to random variation that is inversely related to sample size.[18] In addition, when perfect performance is required to be a top hospital, hospitals with more patients have more opportunities to experience at least 1 failure. As a result, there is an inherent bias against larger centers.

Several methods have been proposed for addressing the issue of small denominators when calculating benchmarks. One simple approach that we examined involves excluding sites with small denominators (eg, <25 eligible cases) when calculating top performance ratings. However, the disparity caused by unequal sample sizes did not disappear when the analysis was restricted to hospitals meeting the reporting threshold of 25 cases. Furthermore, the exclusion of low-volume sites would eliminate at least 32% of US medical centers from consideration as top hospitals for each of the process measures assessed.

An alternative approach to determining benchmarks, known as the ABC method, was also examined in our study. The ABC method has the advantage of reducing the influence of small denominators while using all of the available data in the benchmark calculation rather than simply eliminating hospitals with few eligible patients. In our study, we found that large hospitals were less likely to attain the ABC benchmark, despite the fact that large hospitals, as a group, had higher rates of process measure adherence. Many of the statistical nuances that contribute to this spurious association were addressed in the articles[7,9] that proposed the ABC method. Recognizing the fallibility of hospital classifications based on raw performance scores, these authors[7,9] advocate a nonpunitive approach to quality improvement based on audits and feedback.

Another alternative approach to increasing the reliability of publicly reported quality measures is to estimate performance using hierarchical models. As noted in the "Alternative Benchmark Methods" subsection in the "Results" section, hierarchical models produce shrunken estimates in which the estimate for a single provider is shrunken toward the overall mean for the entire population of providers. This shrinkage property prevented the benchmarks from being artificially inflated by sites with apparently perfect performance based on an extremely small sample size. As a result, in our study, top hospital status was no longer dominated by small hospitals when conventional estimates were replaced by shrunken hierarchical estimates.

However, the "shrinkage" property of hierarchical estimates may introduce a bias in the opposite direction. *Shrinkage* refers to the property that each hospital's hierarchical estimate is adjusted toward the overall mean of all hospitals. The size of this adjustment varies from hospital to hospital and is greater for hospitals with a small sample size. Because of this shrinkage adjustment, it is relatively difficult for a small-volume center to have a point estimate that is extreme enough to exceed 90% of the other hospitals, especially because large hospitals are not adjusted (shrunken) by the same amount. In our study, no hospital with fewer than 10 cases exceeded the hierarchical benchmark for use of ACE inhibitors, aspirin given at arrival, aspirin prescribed at discharge, β-blocker given at arrival, β-blocker prescribed at discharge, PCI administered within 120 minutes, or smoking cessation counseling. Thus, further research may be needed to determine the best method of classifying hospitals and to ensure that centers are compared equitably. It is possible that hospital performance within volume categories can be assessed or that hierarchical models can be modified to directly adjust for hospital volume. Other advanced statistical methods (eg, Bayesian inference, decision theory) were beyond the scope of this study but are known to offer potential advantages when classifying hospitals.[11,19,20]

Although we focused on assessing pitfalls in the identification of top hospitals, the same issues would arise in the identification of poorly performing centers. In general, centers with a small sample size are likely to have extreme results (in either direction), owing to the luckiness or unluckiness of their limited opportunities. We focused on top hospitals because several public reporting and pay-for-performance programs currently reward top providers without explicitly penalizing poor performance.

Beyond public reporting, small and unequal sample sizes also pose challenges when performing data analysis for outcomes research (for example, studies of the association between case volumes and performance on quality measures). Researchers should account for uncertainty when estimating a hospital's performance and use analysis techniques (such as hierarchical models) that partition variance in an appropriate manner.

In addition to hospitals, small case volumes also present challenges when measuring other units, such as nursing homes and practice groups. With the advent of the CMS Physician Quality Reporting Initiative,[21] there is a movement toward performance reporting at the level of individual care providers. The problems described herein will naturally be more acute when data are reported separately for individuals rather than aggregated across individuals within an organization.

In conclusion, reports that ignore the size of the denominator when assessing process performance may unfairly reward or penalize hospitals and mislead consumers. In our study, larger volume hospitals had better aggregate performance, yet were less likely to be identified as top hospitals. Alternative statistical methods that account for small sample sizes may permit a fairer assessment of hospital process performance.

Accepted for Publication: December 3, 2007.

Correspondence: Sean M. O'Brien, PhD, Duke Clinical Research Institute, Box 17969, Durham, NC 27715 (obrie027@mc.duke.edu).

Author Contributions: Dr O'Brien had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. *Study concept and design:* O'Brien, DeLong, and Peterson. *Acquisition of data:* O'Brien and Peterson. *Analysis and interpretation of data:* O'Brien and Peterson. *Drafting of the manuscript:* O'Brien and Peterson. *Critical revision of the manuscript for important intellectual content:* DeLong and Peterson. *Statistical analysis:* O'Brien and DeLong. *Obtained funding:* Peterson. *Administrative, technical, and material support:* Peterson. *Study supervision:* Peterson.

Financial Disclosure: None reported.

REFERENCES

1. Centers for Medicare & Medicaid Services (CMS)/Premier Hospital Quality Incentive Demonstration Project: project overview and findings from year one, 2006. Premier Web site. http://www.premierinc.com/quality-safety/tools-services/p4p/hqi/hqi-whitepaper041306.pdf. Accessed April 16, 2006.
2. Rosenthal MB, Landon BE, Normand SL, Frank RG, Epstein AM. Pay for performance in commercial HMOs. *N Engl J Med.* 2006;355(18):1895-1902.
3. Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? *Ann Intern Med.* 2006;145(4):265-272.
4. Hospital Compare Web site. www.hospitalcompare.hhs.gov. Accessed November 25, 2007.
5. Kiefe CI, Weissman NW, Allison JJ, Farmer R, Weaver M, Williams OD. Identifying achievable benchmarks of care: concepts and methodology. *Int J Qual Health Care.* 1998;10(5):443-447.
6. Hospital Quality Alliance fact sheet. Hospital Quality Alliance Web site. http://www.cms.hhs.gov/HospitalQualityInits/downloads/HospitalHQAFactSheet200512.pdf. Accessed June 28, 2007.
7. Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. *JAMA.* 2001;285(22):2871-2879.
8. Kiefe CI, Woolley TW, Allison JJ, Box JB, Craig AS. Determining benchmarks: a data-driven search for the best achievable performance. *Clin Perform Qual Health Care.* 1994;2(4):190-194.
9. Weissman NW, Allison JJ, Kiefe CI, et al. Achievable benchmarks of care: the ABCs of benchmarking. *J Eval Clin Pract.* 1999;5(3):269-281.
10. Burgess JF Jr, Christiansen CL, Michalak SE, Morris CN. Medical profiling: improving standards and risk adjustments using hierarchical models. *J Health Econ.* 2000;19(3):291-309.
11. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med.* 1997;127(8, pt 2):764-768.
12. Weissman NW, Kiefe CI, Allison J, et al. Achievable Benchmarks of Care (ABC) User Manual. http://main.uab.edu/show.asp?durki=11311. Accessed June 28, 2007.
13. 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.
14. Krumholz HM, Brindis RG, Brush JE, et al; American Heart Association; Quality of Care and Outcomes Research Interdisciplinary Writing Group; Council on Epidemiology and Prevention; Stroke Council; American College of Cardiology Foundation. Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council; endorsed by the American College of Cardiology Foundation. *Circulation.* 2006;113(3):456-462.
15. Canto JG, Every NR, Magid DJ, et al; National Registry of Myocardial Infarction 2 Investigators. The volume of primary angioplasty procedures and survival after acute myocardial infarction. *N Engl J Med.* 2000;342(21):1573-1580.
16. Magid DJ, Calonge BN, Rumsfeld JS, et al. Relation between hospital primary angioplasty volume and mortality for patients with acute MI treated with primary angioplasty vs thrombolytic therapy. *JAMA.* 2000;284(24):3131-3138.
17. Thiemann DR, Coresh J, Oetgen WJ, Powe NR. The association between hospital volume and survival after acute myocardial infarction in elderly patients. *N Engl J Med.* 1999;340(21):1640-1648.
18. Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. *Int J Qual Health Care.* 2001;13(6):481-488.
19. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-care Evaluation.* West Sussex, England: John Wiley &Sons Ltd; 2004.
20. Austin PC, Anderson GM. Optimal statistical decisions for hospital report cards. *Med Decis Making.* 2005;25(1):11-19.
21. Physician quality reporting initiative. Centers for Medical & Medicaid Services Web site. http://www.cms.hhs.gov/pqri. Accessed April 16, 2008.