© Published 2014. This article is a U.S. Government work and is in the public domain in the U.S.A.

DOI: 10.1111/1475-6773.12149

METHODS ARTICLE

Statistical Benchmarks for Health Care Provider Performance Assessment: A Comparison of Standard Approaches to a Hierarchical Bayesian Histogram-Based Method

Susan M. Paddock

Objective. Examine how widely used statistical benchmarks of health care provider performance compare with histogram-based statistical benchmarks obtained via hierarchical Bayesian modeling.

Data Sources. Publicly available data from 3,240 hospitals during April 2009–March 2010 on two process-of-care measures reported on the Medicare Hospital Compare website.

Study Design. Secondary data analyses of two process-of-care measures comparing statistical benchmark estimates and threshold exceedance determinations under various combinations of hospital performance measure estimates and benchmarking approaches.

Principal Findings. Statistical benchmarking approaches for determining top 10 percent performance varied with respect to which hospitals exceeded the performance benchmark; such differences were not found at the 50 percent threshold. Benchmarks derived from the histogram of provider performance under hierarchical Bayesian modeling provide a compromise between benchmarks based on direct (raw) estimates, which are overdispersed relative to the true distribution of provider performance and prone to high variance for small providers, and posterior mean provider performance, for which over-shrinkage and under-dispersion relative to the true provider performance distribution is a concern.

Conclusions. Given the rewards and penalties associated with characterizing top performance, the ability of statistical benchmarks to summarize key features of the provider performance distribution should be examined.

Key Words. Bayesian statistics, hierarchical model, provider profiling, public reporting, statistical benchmark

The development and public reporting of performance standards are central to large-scale efforts to monitor and improve health care provider quality (McNamara 2006). Performance standards are also important for pay-for-performance programs (Rosenthal et al. 2005). These initiatives frequently involve characterizing provider performance using quality measures (QMs). For a given provider, a QM may summarize clinical processes of care, patient outcomes, or patient experiences with care. One example of such a data collection effort is the Healthcare Effectiveness Data and Information Set (HEDIS), which contains data on 76 QMs collected across five domains of care from over 90 percent of U.S. health care plans. The National Committee for Quality Assurance (NCQA) requires health plans to submit HEDIS measures for its accreditation process. For the HEDIS portion of the accreditation score, NCQA compares the plan's results to a national benchmark (the 90th percentile of national results) and to regional and national performance thresholds (the 75th, 50th, and 25th percentiles). Similarly, the Medicare Hospital Compare Web site reports the performance of hospitals meeting QMs covering process of care, patient experiences with care, and outcomes. The Medicare Hospital Value-Based Purchasing (VBP) program implemented in fiscal year (FY) 2013 provides additional payments to hospitals with relatively strong performance on a collection of QMs. An important component of determining the hospital bonus amounts is to compare each QM for a given hospital to a performance standard, which is defined as the 50th percentiles of provider performance on that QM. A hospital must exceed this threshold to receive any points for the QM in the bonus payment formula, with the maximum number of points given to those hospitals that exceed the performance benchmark, which is the mean performance of the top 10 percent of providers.

Characterizing top performance using percentile cutoffs has also been advocated for smaller, local continuous quality improvement efforts. For example, the Achievable Benchmarks of Care (ABC) (Kiefe et al. 2001) approach defines the performance of the top 10 percent of providers as representing a "realistic standard of excellence" due to having been attained by some providers. Physicians who were randomized to receive feedback about their performance using the ABC method significantly improved their performance relative to others. Although the ABC method was initially developed for application to continuous quality improvement efforts, it has been employed elsewhere as a way to develop domain-wide performance

benchmarks in areas such as mental health (Hermann et al. 2006) and primary care (Wessell et al. 2008).

Despite the importance of statistical benchmarks to local and national quality improvement efforts, there has been very little exploration of the methodology used for their development. One exception is an analysis of data downloaded from the Hospital Compare Web site that found ABC and methods similar to the one used under the Medicare Hospital VBP program that identify the top 10 percent of providers tend to identify smaller, lower-volume hospitals as being in the top 10 percent, despite exploratory data analyses showing that higher-volume hospitals actually have better performance (O'Brien, DeLong, and Peterson 2008). One explanation provided for this finding is that direct (or raw) estimates of provider performance, such as the observed proportion of times a QM is met for a provider (numerator) out of all possible opportunities (denominator), have high variance, particularly for small providers. In response, O'Brien, DeLong, and Peterson (2008) estimated an alternative statistical benchmark of top 10 percent performance as the 90th percentile of posterior means of provider performance derived from a two-stage Bayesian hierarchical model. This model has gained popularity for profiling health care provider performance due to the greater stability of posterior means versus raw means, particularly for small providers, when the focus is on obtaining an estimate of true provider-specific performance (Christiansen and Morris 1997; Normand, Glickman, and Gatsonis 1997; Burgess et al. 2000; Landrum, Bronskill, and Normand 2000). This stability results from each provider's posterior mean being a shrinkage estimator, which is a weighted average of the provider's raw mean performance estimate and the overall (or "grand") mean of the performance across all hospitals. Relatively greater weight is given to the overall mean in the hospital's posterior mean when its raw mean has large variance, while less weight is given to the overall mean when a hospital's raw performance estimate has relatively low variance. However, relative to the raw means approach, O'Brien, DeLong, and Peterson (2008) found that the relationship between hospital size and performance was reversed using posterior means, raising concerns that using posterior means over-compensates for small hospital size (higher variance) by weighting the grand mean so heavily for small providers that it becomes very unlikely for small providers to exceed the 90th percentile. Similar concerns have been raised when the goal is to estimate provider-specific performance for low-volume providers (Mukamel et al. 2010; Silber et al. 2010).

Another explanation for the findings of O'Brien, DeLong, and Peterson (2008) that has implications regardless of hospital size is that estimating the

90th percentile of performance and identifying which providers exceed that performance are questions that pertain to the distribution of performance across all providers rather than summaries of hospital-specific performance (Paddock and Louis 2011). The posterior means are optimal under squared-error loss for estimating performance for individual providers but are under-dispersed relative to the distribution of the true, underlying provider parameters (Shen and Louis 1998). In addition, direct estimates of provider performance—that is, those calculated as simply the performance for each provider separately, rather than pooling data using hierarchical models—are over-dispersed for estimating the distribution of provider performance (Shen and Louis 1998). The histogram, or empirical distribution function, of provider-specific performance provides the optimal estimate of the distribution of provider-specific performance under integrated squared error loss. Figure 1 illustrates this difference.¹ The histogram of the posterior means is underdispersed relative to the true histogram of provider performance, which is denoted by the solid curve (Figure 1, left panel). The raw estimates (Figure 1, middle panel) are over-dispersed relative to the true histogram. The estimated histogram (Figure 1, right panel) correctly characterizes the true histogram of

Figure 1: Illustration Using Simulated Data of the True Distribution of Provider-Specific Performance (solid line) versus Distribution (gray) Based on (a) Posterior Means, (b) Raw Estimates, and (c) Histogram Method

provider performance. Selecting the 90th percentile of provider performance by examining the right-most side of each histogram reveals that the 90th percentile of the posterior means would be too small, while that of the raw mean estimates would be too large.

Despite its appropriateness for estimating the distribution of performance for a population of providers, histogram-based statistical benchmarking has yet to be compared to widely used methods. In this article, their performance relative to other more widely used statistical benchmarking methods is examined to demonstrate whether and how their performance differs, both in terms of setting a performance benchmark and also in terms of which providers are identified as exceeding the performance benchmark. The stability of conclusions drawn about provider performance when different estimation procedures for provider-specific performance are compared with statistical benchmarks is also examined. These methods are illustrated using data on two processes-of-care measures reported on the Medicare Hospital Compare Web site and used in the Medicare Hospital VBP program.

METHODS

Data

Data come from the Centers for Medicare and Medicaid Services' Hospital Compare Hospital Quality Initiatives website, archived December 2010 (data covering the second quarter 2009 through the first quarter 2010). Data collection for Hospital Compare is retrospective and includes data elements from administrative data and medical record documents. Data collected from general acute care facilities are examined here. All facilities that had at least one eligible case for the QMs examined here are included. As the number of eligible cases (denominator) and percent of cases meeting a QM are provided in the dataset for each hospital, the numerator was estimated as the denominator multiplied by the percent of cases meeting the QM, rounded to the nearest integer. The effect of benchmark methodology is examined by examining two representative QMs, AMI-8a (primary percutaneous coronary interventions received within 90 minutes of hospital arrival) and PN-6 (initial antibiotic selection for community acquired pneumonia in immunocompetent patient). Descriptive statistics for each of these measures are provided in Table 1. Data are available on AMI-8a for 1,546 hospitals, although Hospital Compare data are available for most QMs for about 3,200 hospitals, as is the case for PN-6. The mean performances of both AMI-8a and PN-6 are 85 and 90 percent,

Table 1: Descriptive Statistics on the Two Quality Measures

	<i>Number of Hospitals</i>	<i>Mean Performance (%)</i>	<i>Percentiles (10th, 50th, 90th)</i>	<i>Hospital Denominator Size (n) Percentiles Mean n (10th, 90th)</i>
AMI-8a: Primary percutaneous coronary interventions (PCI) received within 90 minutes of hospital arrival	1,546	85%	(66, 90, 100)	36 (9, 66)
PN-6: Initial antibiotic selection for community acquired pneumonia in immunocompetent patient	3,240	90%	(82, 92, 98)	106 (35, 193)

respectively, which is commensurate with all but one of the other performance measures. There is more spread in the percentiles of the hospital raw means, each estimated as numerator divided by denominator, for AMI-8a than PN-6, which is expected given the relatively small average denominator size per hospital of 36 for AMI-8a versus 106 for PN-6. Some hospitals have relatively small (less than 30) denominators for AMI-8a, whereas the minimum denominator for PN-6 is 35. The raw estimates for AMI-8a are thus more variable than those for PN-6.

Analytic Approach

In this article, statistical benchmarking techniques are divided into two sets: approaches that are commonly used in practice and those based on hierarchical modeling. For clarity of exposition, the paper focuses on the canonical case of a QM for provider j being defined by its numerator r_j , which is the number of occasions in which the care characterized by the QM was delivered, and its denominator n_j , which is the number of opportunities provider j had to deliver the QM. However, the ideas set forth apply to other QM data types.

Widely Used Statistical Benchmarking Approaches. HEDIS Approach (RAW): The statistical benchmark is the 90th percentile of the observed performance of all providers. Performance for each provider j is estimated as the numerator divided by the denominator of the performance measure: $EST-RAW_j = r_j/n_j$.

Modified HEDIS Approach (RAW-10): Similar to EST-RAW, the statistical benchmark is the 90th percentile of the r_j/n_j 's, but only for hospitals with at

least 10 observations, an approach used by the Medicare Hospital VBP program (Department of Health and Human Services 2011). The goal is to reduce the influence of low-volume hospitals on the benchmark given concerns about the stability of estimates from low-volume hospitals.

Achievable Benchmark of Care Approach (Kiefe et al. 2001): Recognizing that very low-volume, high-variance providers are prone to having extreme observed performance (e.g., EST-RAW = 0 or 1) due to chance, the approach of Kiefe et al. (2001) reduces their influence on statistical benchmarks through the use of a Bayesian estimator. The adjusted performance fraction (APF) modifies EST-RAW by adding 1 to its numerator and 2 to its denominator: $APF_j = (1 + r_j)/(2 + n_j)$. This is equivalent to assuming hospital j 's performance data follow a binomial sampling model with a parameter, p_j , that characterizes the true performance of hospital j , and then placing a uniform prior distribution on p_j , which results in smoothing estimates for low-volume hospitals away from 0 or 1 (Gelman et al. 2003). The ABC benchmark is estimated by ranking all providers from high to low on APF_j until the top providers identified include 10 percent of all patients in the population. The ABC benchmark is set equal to the number of eligible cases meeting the QM divided by the number of eligible cases pooled across these top providers.

Medicare Hospital Compare Approach (COMP): Similarly to the modified HEDIS approach (RAW-10), the COMP statistical benchmark is estimated by first ranking from highest to lowest the raw mean hospital performance for those hospitals with denominators of at least 10, and then selecting the top 10 percent of hospitals based on this ranking. The COMP benchmark differs from the RAW-10 benchmark in that it is the average of the raw mean performances of this top 10 percent.

Statistical Benchmarks Based on Bayesian Hierarchical Modeling. The first step in developing a statistical benchmark from a Bayesian hierarchical model is to specify the model. Since all QMs examined in this article are dichotomous at the patient level, a two-stage hierarchical Bayesian logistic regression model is assumed for the data and used to illustrate the approach. The first stage of the model is to specify the data for each hospital $k = 1, \dots, K$, follow a binomial sampling model: $r_k | n_k, p_k \sim \text{Binomial}(n_k, \theta_k)$, where θ_k is a parameter that characterizes hospital performance and θ_k represents the true but unknown proportion of the time that hospital k meets the QM. The second stage of the model involves linking hospital-specific random effects, b_k , to θ_k which is done here by selecting a logistic link function: $\text{logit}(\theta_k) = b_k$. The hospital-spe-

cific random effects, b_k , are assumed to follow a common distribution G . In the analyses presented here, G is chosen to be Gaussian. The choices of G being Gaussian and of a logistic link function are typical given that most widely used hierarchical modeling software packages only allow for G to be Gaussian (Li et al. 2011). However, other choices for the link function are possible, and G could be chosen differently, such as specifying a Beta distribution (Adams 2009) or modeling G nonparametrically (Paddock et al. 2006). This two-level model structure facilitates a borrowing of strength among the provider parameters, the θ_k 's, so that each θ_k is estimated as a function of the provider's raw estimate (i.e., r_k/n_k) and the mean provider performance, $\pi = \exp(\mu)/(1 + \exp(\mu))$; θ_k is weighted more heavily toward r_k/n_k if hospital k has lower variation in performance, but more heavily weights π for hospitals with relatively noisy variance in performance.²

Posterior Mean Approach (O'Brien, DeLong, and Peterson 2008): One may obtain estimates of the mean performance for each hospital k as the posterior mean of θ_k , or EST-PM $_k$. It may be estimated as the average of all saved MCMC draws. The posterior mean-based statistical benchmark equals the 90th percentile of the EST-PM $_k$'s. The relative increase in stability of performance estimates, particularly for high variance or low-volume hospitals, under the Bayesian hierarchical model relies on correctly specifying the distribution of hospital performance. Hospital characteristics, Z_k , that are important for determining performance may be added to the model by replacing logit (θ_k) = b_k with $\text{logit}(\theta_k) = b_k + aZ_k$. This would allow, for example, modeling an association between hospital volume and performance if such a relationship were present in the data, resulting in performance for low (high) volume providers not being shrunk not toward the overall mean but rather reflecting the mean for low (high) volume providers (Mukamel et al. 2010; Silber et al. 2010).

Histogram Approach (Paddock and Louis 2011): Unlike posterior means or raw means, the histogram of the K provider performance parameters, $\theta_1, \dots, \theta_K$, is useful for making inferences about the distribution of performance of the K hospitals, such as obtaining percentiles (Shen and Louis 1998). The histogram is summarized using MCMC output as follows (Conlon and Louis 1999): Set U_k equal to the q th quantile of all of the MCMC samples of ($\theta_1, \dots, \theta_K$) for $q = (2k - 1)/(2K)$, ($k = 1, \dots, K$). The p th percentile of provider performance is the histogram-based statistical benchmark, *HIST*, which is estimated as the U_k corresponding to the smallest q that exceeds p . The U_k 's may also be used to derive a hospital-specific performance estimate that is less affected by the shrinkage in the hierarchical model than is the posterior mean:

$\text{EST-HIST}_k = U_{R_k}$, where R_k is the posterior integer rank of θ_k (Shen and Louis 1998).

RESULTS

Table 1 provides descriptive statistics on the two QMs examined in this article and are typical of the process measure QMs reported on the Medicare Hospital Compare Web site. The range of performance is greater for AMI-8a (the 10th and 90th percentiles are 66 and 100 percent, respectively) than for PN-6. The 90th percentile of raw performance is at or near 100 percent for both measures, which is typical of the process measures in Hospital Compare. Denominators for AMI-8a tend to be relatively small, with the mean denominator 36; in contrast, denominators for PN-6 are more moderate, with a mean of 106.

Each column of Table 2 represents one of the six approaches described above to obtain a statistical benchmark of provider performance, and each row provides the statistical benchmark estimates for AMI-8a and PN-6. The benchmarks derived from posterior means (PM) are lower than the others, while those based on RAW, RAW-10, ABC, and COMP are among the highest and are very similar in value. The HIST benchmark provides a compromise between PM and the other methods, which is expected given the comparison across methods illustrated in Figure 1 and the similarities among RAW, RAW-10, ABC, and COMP. Even though the benchmarks all appear qualitatively similar given they are all relatively high values, their range of 3–4 percentage points suggests that hospitals could differ considerably in meeting some but not all of the benchmarks, given the similarly highly centered distributions of hospital performances shown in Table 1.

The exceedance of a benchmark depends not only on how the benchmark is estimated but also on how the hospital-specific performance estimate is derived. Thus, Table 3 displays the number of statistical benchmarks exceeded out of the six described above and shows how this varies depending on whether the hospital-specific estimate is the raw estimate, PM, or the

Table 2: Statistical Benchmark Estimates

	<i>RAW</i>	<i>RAW-10</i>	<i>ABC</i>	<i>COMP</i>	<i>PM</i>	<i>HIST</i>	<i>Range</i>
AMI-8a	1.0000	1.0000	0.9962	1.0000	0.9599	0.9738	0.0401
PN-6	0.9800	0.9787	0.9887	0.9927	0.9619	0.9697	0.0308

Table 3: Number of Benchmarks Exceeded When Using RAW, HIST, or PM to Characterize Individual Provider Performance

Provider performance estimate:	Number of Benchmarks Exceeded						
	0	1	2	3	4	5	6
AMI-8a							
RAW <i>N</i>	1,135	98	72	0	0	0	241
%	73	6	5	0	0	0	16
HIST <i>N</i>	1,248	143	153	2	0	0	0
%	81	9	10	0	0	0	0
PM <i>N</i>	1,391	132	23	0	0	0	0
%	90	9	1	0	0	0	0
PN-6							
RAW <i>N</i>	2,575	149	163	27	122	41	163
%	79	5	5	1	4	1	5
HIST <i>N</i>	2,708	208	189	21	95	15	4
%	84	6	6	1	3	0	0
PM <i>N</i>	2,916	187	112	4	19	2	0
%	90	6	3	0	1	0	0

triple-goal estimate (HIST). For both measures, RAW hospital estimates are more likely to exceed at least one benchmark and are also more likely to exceed all benchmarks. Under the RAW hospital-specific estimate, a greater percentage of hospitals exceeded at least one benchmark on the AMI-8a measure (27 percent) versus the PN-6 measure (21 percent); this is explained by hospitals having smaller denominators for AMI-8a and thus tending to have higher variance raw hospital performance estimates for AMI-8a that are more subject to being relatively extreme. For both AMI-8a and PN-6, the HIST (triple-goal) estimates provide a compromise in terms of hospitals being flagged as exceeding at least one performance threshold. Using either the HIST and PM estimates result in fewer hospitals exceeding multiple performance thresholds than RAW estimates.

Table 4 shows the percent agreement in whether hospitals simultaneously exceed each pair of statistical benchmarks that denote “top 10 percent” performance. Since the determination of whether a hospital exceeds a benchmark depends on the benchmark and on the hospital-specific estimate, the percent agreement matrix is reported for each type of hospital-specific estimate (for AMI-8a, Table 4a–c, respectively; for PN-6, Table 4d–f, respectively). The row and column names denote the two statistical benchmarks being compared for each estimate. For example, in Table 4b, the upper right-hand estimate of 99 percent in the RAW row and the HIST column means that

Table 4: Number of Top 10% Facilities (Yes vs. No) Using Statistical Benchmarks When Estimating Individual Provider Performance Using (a) RAW-EST, (b) HIST-EST, (c) PM-EST

<i>Benchmark</i>	<i>RAW-10, %</i>	<i>ABC, %</i>	<i>COMP, %</i>	<i>PM, %</i>	<i>HIST, %</i>
Exceedance agreement for AMI-8a					
(a) EST-RAW					
RAW	100	100	100	89	95
RAW-10		100	100	89	95
ABC			100	89	95
COMP				89	95
PM					100
(b) EST-PM					
RAW	100	100	100	90	99
RAW-10		100	100	90	99
ABC			100	90	99
COMP				90	99
PM					100
(c) EST-HIST					
RAW	100	100	100	81	90
RAW-10		100	100	81	90
ABC			100	81	90
COMP				81	90
PM					100
Exceedance agreement for PN-6					
(d) EST-RAW					
RAW	99	96	95	90	94
RAW-10		95	94	90	95
ABC			99	86	90
COMP				85	89
PM					100
(e) EST-PM					
RAW	100	99	99	91	96
RAW-10		99	99	91	97
ABC			100	90	96
COMP				90	96
PM					100
(f) EST-HIST					
RAW	99	97	97	87	94
RAW-10		96	96	88	94
ABC			100	84	91
COMP				84	90
PM					100

when hospital-specific performance is estimated using posterior means (EST-PM), the hospital performance exceeds both the RAW and HIST statistical benchmarks 99 percent of the time. The RAW, RAW-10, ABC, and

COMP benchmarks have the highest levels of agreement for both AMI-8a and PN-6, which is explained by the fact they are estimated very similarly and relatively similar values (Table 1). The PM and HIST benchmarks agree 100 percent of the time as well. For both AMI-8a and PN-6, the nonhierarchical model-based benchmarks (RAW, RAW-10, ABC, and COMP) have lower percent agreements with PM than with HIST. Regardless of the hospital-specific estimate used, the HIST-based benchmark provides a compromise between the PM, which is under-dispersed for the true provider performance distribution, and the nonhierarchical model-based benchmarks, which are prone to flagging high-variance providers as exceeding the benchmark. The percent agreement ranges from 81 to 100 percent across all combinations of hospital-specific estimates and statistical benchmarks. A table similar to Table 4 for estimating a benchmark at the 50th percentile of performance is not included here, but it shows the range of percent agreement to be 95–100 percent for AMI-8a and 99–100 percent for PN-6. This relatively consistent performance is expected given that relatively moderate percentiles such as the median may be estimated more precisely than more extreme percentiles for a given dataset (Garsd et al. 1983).

CONCLUSION

Despite the policy relevance of statistical benchmarking for characterizing top performance in public reporting and setting targets for pay-for-performance programs, very little attention has been devoted to examining methods for setting statistical benchmarks. Given the direct monetary awards and penalties (Rosenthal et al. 2006), indirect effects of evaluating and reporting provider performance, such as increased or decreased patient referrals (Werner and Asch 2005), and limited tolerance for misclassification risk among some consumers of publicly reported provider reports (Davis, Hibbard, and Milstein 2007), it is important for statistical benchmarks to best characterize the policy-relevant inferential target, which is the distribution of provider performance. Contributions of this article include conducting such an examination, introducing the histogram benchmarking approach to the health services research community and explaining how it is distinct from widely used benchmarking approaches, and highlighting the limitations of standard benchmarking approaches. The work presented here complements the statistical and health services research literature on hierarchical modeling methods for provider performance measurement, which almost exclusively focuses on estimating

performance or ranks for individual providers rather than examining the distribution of performance in the provider population.

For the denominator sizes found in the Medicare Hospital Compare data, the choice of statistical benchmarking approach used to determine top 10 percent performance results in some discrepancies as to which hospitals exceed the performance benchmark. None of the benchmarks examined here align perfectly with the histogram approach, which has the desirable statistical property of being optimal under integrated squared error loss for percentile estimation. As hypothesized, the histogram approach mitigates problems with raw estimates found by O'Brien, DeLong, and Peterson (2008)—which are over-dispersed for the distribution of provider performance—and posterior means—which are under-dispersed for the distribution of provider performance—by offering a compromise between these two extremes.

A limitation of the histogram approach is its greater conceptual and computational complexity relative to the more widely used RAW, ABC, or COMP approaches. However, health services researchers have been actively applying similarly complex analytic approaches to mitigate the difficulties posted by high-variance raw estimates, such as statistical testing-based approaches (Adams et al. 2010) and hierarchical modeling to separate true performance from measurement error (COPSS-CMS White Paper Committee 2012). Consideration of the histogram approach is a natural progression of this line of work. At a minimum, this article illustrates how one might evaluate the usefulness of simple, yet nonoptimal, benchmarking approaches by comparing them to one obtained from an approach that is optimal with respect to an appropriate loss function for targeting the true distribution of provider performance.

Given the reliance on Bayesian theory, some might be concerned about the sensitivity to prior distributional assumptions underlying the histogram and posterior mean approaches. One aspect is sensitivity to prior assumptions, given the model specification. This can be mitigated by carefully selecting prior distributions, particularly for variance components when very few hospitals are in the population to be profiled (Gelman 2006), and employing sensitivity analyses to confirm the stability of inferences to prior distributional assumptions by fitting the model with several different plausible yet relatively uninformative prior distributions. Other potential areas of concern are the correctness of the hierarchical model specification with respect to parametric form of the hospital, or random effects, distribution (Lin et al. 2009) and to the assumption that all hospitals come from a common population distribution (Mukamel et al. 2010). To

address the former, one may conduct nonparametric or semiparametric analyses (Austin 2009) to protect against model mis-specification, although it is important to note that Bayesian nonparametric models are not assumption-free (Paddock et al. 2006). To address the latter, one may shrink toward the mean for a relevant set of hospitals by including relevant hospital characteristics into the model, such as shrinking low-volume providers toward the mean of low-volume hospitals (Silber et al. 2010). Policy makers would need to decide whether low-volume hospitals as a group should have their estimates shrunk to the mean of low-volume hospitals or, as is done in the model used in this analysis, have all hospitals' estimates shrunk to the mean for all hospitals. Given the difficulties posed by estimating low-volume provider performance, alternative approaches for stabilizing their estimates might be considered before using them as inputs to statistical benchmark calculation, such as basing hospital performance estimates on multiple years of data when the resulting pooled sample leads to meaningful improvements in the precision of provider estimates (Elliott et al. 2009; Lin et al. 2009). While attractive in terms of increasing statistical precision, designers of public reports would need to assess how many years of data to use and whether the use of older information would be acceptable to stakeholders.

The objective of statistical benchmarking is to improve provider performance by setting high yet realistic standards. Benchmarking requires the examination of the distribution of performance for a population, or ensemble, of providers. This contrasts with the goal of estimating performance for any given individual provider. The method of analysis should be guided by the choice of an appropriate loss function given the analytic goal (Shen and Louis 1998). Thus, other methods should be considered for alternative goals, such as identifying individual providers having outlier or extreme (divergent) performance (Ohlssen, Sharples, and Spiegelhalter 2007; Jones & Spiegelhalter 2011; Kalbfleisch and Wolfe 2013).

The analysis presented here highlights differences in how specific providers might be recognized and rewarded in statistical benchmarking efforts. The implications of using alternative statistical benchmarking approaches might vary for different types of providers and different quality measures; for example, benchmarks would be more likely to agree across methods when provider sample sizes are large. In the context of the Medicare Hospital VBP program, future work could examine whether and how different benchmarking approaches affect the bonus payments ultimately received by hospitals, given the multiple measures that must be benchmarked and the assignment

and aggregation of measure-specific scores based on how hospitals' quality measures compare to the VBP's performance thresholds and achievement benchmarks.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This project was supported by grant number R21HS021860 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

Disclosures: None.

Disclaimers: None.

NOTES

1. To demonstrate the differences across estimators when the true distribution of hospital performance is known, data are simulated for $K = 100$ hospitals from a two-stage hierarchical models as follows, using the notation from Statistical Benchmarks based on Bayesian Hierarchical Modeling Section: for each of $k = 1, \dots, K = 100$ hospitals, θ_k are independently drawn from a $N(0,1)$ distribution, so the true hospital performance distribution is $N(0,1)$, as depicted by the solid black lines in Figure 1. Then, $\mathbf{y}_k$ are simulated from a $N(\theta_k, \sigma_k^2)$, where $\mathbf{y}_k$ represents the observations for hospital k and σ_k^2 is the variance of hospital k 's observations, and the σ_k^2 's are chosen so their geometric mean is 1 and $\max\{\sigma_k^2\}/\min\{\sigma_k^2\} = 100$, resulting in a mix of high, medium, and low variance hospitals in the simulated data.
2. The prior distributions for the unknown parameters of G —its mean, μ , and variance, τ^2 —are $p(\mu) \propto 1$ and $\tau \sim \text{Uniform}(0,A)$, where $A = 100$. A is selected to be large enough to cover all plausible values of the standard deviation of the b_k 's (Gelman 2006). Inferences can be particularly sensitive to the prior distribution when τ is near zero or there are very few (<10) hospitals. Even though the data presented in this article have neither of these features, the results presented here were robust under alternative priors: $A = 5$ and $\tau^{-2} \sim \text{Gamma}(0.001, 0.001)$, where the a priori expected value of τ^{-2} equals 1. Marginal posterior distributions of all unknown model parameters were simulated using Markov Chain Monte Carlo (MCMC) as implemented in WinBUGS software (Lunn et al. 2000). Two Markov chains were initiated at overdispersed starting values and each chain ran for 1,000 iterations to allow for the Markov chains to converge to the posterior distribution. These initial samples were discarded, and statistical inferences are based on the subsequent 1,000 iterations from the two chains. Convergence of the MCMC algorithm was

monitored through visual inspection of MCMC trajectories and the Gelman-Rubin convergence diagnostic (Gelman and Rubin 1992).

REFERENCES

- Adams, J. L. 2009. *The Reliability of Provider Profiling: A Tutorial*. TR-653-NCQA. Santa Monica, CA: RAND Corporation.
- Adams, J. L., E. A. McGlynn, J. W. Thomas, and A. Mehrotra. 2010. "Incorporating Statistical Uncertainty in the Use of Physician Cost Profiles." *BMC Health Services Research* 10 (57): 57.
- Austin, P. C. 2009. "Are (the Log-Odds of) Hospital Mortality Rates Normally Distributed? Implications for Studying Variations in Outcomes of Medical Care." *Journal of Evaluation in Clinical Practice* 15 (3): 514–23. doi:10.1111/j.1365-2753.2008.01053.x.
- Burgess, J. F., Jr., C. L. Christiansen, S. E. Michalak, and C. N. Morris. 2000. "Medical Profiling: Improving Standards and Risk Adjustments Using Hierarchical Models." *Journal of Health Economics* 19 (3): 291–309.
- Christiansen, C. L., and C. N. Morris. 1997. "Improving the Statistical Approach to Health Care Provider Profiling." *Annals of Internal Medicine* 127 (8): 764–8.
- Conlon, E. M., and T. A. Louis. 1999. Addressing Multiple Goals in Evaluating Region-Specific Risk Using Bayesian Methods. In *Disease Mapping and Risk Assessment for Public Health*, edited by A. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J. F. Viel, and R. Bertollini pp. 31–47. Chichester, UK: Wiley.
- COPSS-CMS White Paper Committee. 2012. Statistical Issues in Assessing Hospital Performance. Retrieved November 6, 2012, from <http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf>
- Davis, M. M., J. H. Hibbard, and A. Milstein. 2007. *Issue Brief: Consumer Tolerance for Inaccuracy in Physician Performance Ratings: One Size Fits None*. Washington, DC: Center for Studying Health System Change.
- Department of Health and Human Services, Centers for Medicare & Medicaid Services. 2011. "Medicare Program; Hospital Inpatient Value-Based Purchasing Program." *Federal Register* 76 (88): 26490–547.
- Elliott, M. N., D. F. McCaffrey, B. K. Finch, D. J. Klein, N. Orr, M. K. Beckett, and N. Lurie. 2009. "Improving Disparity Estimates for Rare Racial/Ethnic Groups with Trend Estimation and Kalman Filtering: An Application to the National Health Interview Survey." *Health Services Research* 44 (5 Pt 1): 1622–39. doi: 10.1111/j.1475-6773.2009.01000.x.
- Garsd, A., G. E. Ford, G. O. Waring, 3rd, and L. S. Rosenblatt. 1983. "Sample Size for Estimating the Quantiles of Endothelial Cell-Area Distribution." *Biometrics* 39 (2): 385–94.
- Gelman, A. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1 (3): 515–33.

- Gelman, A., and D. B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7 (4): 457–72.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin 2003. *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.
- Jones, H. E., and D. J. Spiegelhalter. 2011. "The Identification of 'Unusual' Health-Care Providers from a Hierarchical Model." *American Statistician* 65 (3): 154–63.
- Hermann, R. C., J. A. Chan, S. E. Provost, and W. T. Chiu. 2006. "Statistical Benchmarks for Process Measures of Quality of Care for Mental and Substance Use Disorders." *Psychiatric Services* 57 (10): 1461–7. doi:DOI 10.1176/appi.ps.57.10.1461.
- Kalbfleisch, J. D., and R. A. Wolfe. 2013. "On Monitoring Outcomes of Medical Providers." *Statistics in Biosciences* 5: 286–302. doi:10.1007/s12561-013-9093-x.
- Kiefe, C. I., J. J. Allison, O. D. Williams, S. D. Person, M. T. Weaver, and N. W. Weissman. 2001. "Improving Quality Improvement Using Achievable Benchmarks for Physician Feedback." *Journal of the American Medical Association* 285 (22): 2871–9. doi:10.1001/jama.285.22.2871.
- Landrum, M. B., S. E. Bronskill, and S.-L. T. Normand. 2000. "Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers." *Health Services and Outcomes Research Methodology* 1 (1): 23–47. doi:10.1023/a:1010093701870.
- Li, B. Y., H. F. Lingsma, E. W. Steyerberg, and E. Lesaffre. 2011. "Logistic Random Effects Regression Models: A Comparison of Statistical Packages for Binary and Ordinal Outcomes." *BMC Medical Research Methodology* 11: Art 77. doi:10.1186/1471-2288-11-77.
- Lin, R., T. Louis, S. Paddock, and G. Ridgeway. 2009. "Ranking USRDS Provider Specific SMRS from 1998–2001." *Health Services and Outcomes Research Methodology* 9 (1): 22–38.
- Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. "WinBUGS – A Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing* 10 (4): 325–37.
- McNamara, P. 2006. "Provider-Specific Report Cards: A Tool for Health Sector Accountability in Developing Countries." *Health Policy and Planning* 21 (2): 101–9. doi:10.1093/heapol/czj009.
- Mukamel, D. B., L. G. Glance, A. W. Dick, and T. M. Osler. 2010. "Measuring Quality for Public Reporting of Health Provider Quality: Making It Meaningful to Patients." *American Journal of Public Health* 100 (2): 264–9. doi:10.2105/AJPH.2008.153759.
- Normand, S.-L. T., M. E. Glickman, and C. A. Gatsonis. 1997. "Statistical Methods for Profiling Providers of Medical Care: Issues and Applications." *Journal of the American Statistical Association* 92 (439): 803–14.
- O'Brien, S. M., E. R. DeLong, and E. D. Peterson. 2008. "Impact of Case Volume on Hospital Performance Assessment." *Archives of Internal Medicine* 168 (2): 1277–84.
- Ohlssen, D., L. D. Sharples, and D. J. Spiegelhalter. 2007. "A Hierarchical Modelling Framework for Identifying Unusual Performance in Health Care Providers." *Journal of the Royal Statistical Society, Series A* 170 (4): 865–90.

- Paddock, S. M., and T. A. Louis. 2011. "Percentile-Based Empirical Distribution Function Estimates for Performance Evaluation of Healthcare Providers." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60 (4): 575–89. doi:10.1111/j.1467-9876.2010.00760.x.
- Paddock, S. M., G. Ridgeway, R. Lin, and T. A. Louis. 2006. "Flexible Distributions for Triple-Goal Estimates in Two-Stage Hierarchical Models." *Computational Statistics & Data Analysis* 50 (11): 3243–62. doi:10.1016/j.csda.2005.05.008.
- Rosenthal, M. B., R. G. Frank, Z. Li, and A. M. Epstein. 2005. "Early Experience with Pay-for-Performance: From Concept to Practice." *Journal of the American Medical Association* 294 (14): 1788–93. doi:10.1001/jama.294.14.1788.
- Rosenthal, M. B., B. E. Landon, S. L. Normand, R. G. Frank, and A. M. Epstein. 2006. "Pay for Performance in Commercial HMOS." *The New England Journal of Medicine* 355 (18): 1895–902. doi:10.1056/NEJMsa063682.
- Shen, W., and T. A. Louis. 1998. "Triple-Goal Estimates in Two-Stage Hierarchical Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60 (2): 455.
- Silber, J. H., P. R. Rosenbaum, T. J. Brachet, R. N. Ross, L. J. Bressler, O. Even-Shoshan, and K. G. Volpp. 2010. "The Hospital Compare Mortality Model and the Volume-Outcome Relationship." *Health Services Research* 45 (5 Pt 1): 1148–67. doi:10.1111/j.1475-6773.2010.01130.x.
- Werner, R. M., and D. A. Asch. 2005. "The Unintended Consequences of Publicly Reporting Quality Information." *Journal of the American Medical Association* 293 (10): 1239–44.
- Wessell, A. M., H. A. Liska, P. J. Nietert, R. G. Jenkins, L. S. Nemeth, and S. Ornstein. 2008. "Achievable Benchmarks of Care for Primary Care Quality Indicators in a Practice-Based Research Network." *American Journal of Medical Quality* 23 (1): 39–46. doi:10.1177/1062860607310774.