Estimating a Composite Measure of Hospital Quality From the Hospital Compare Database

Differences When Using a Bayesian Hierarchical Latent Variable Model Versus Denominator-Based Weights

Michael Shwartz, PhD,† Justin Ren, PhD,† Erol A. Peköz, PhD,†* Xin Wang, PhD,‡ Alan B. Cohen, ScD,§†* and Joseph D. Restuccia, DrPH†**

Background: A single composite measure calculated from individual quality indicators (QIs) is a useful measure of hospital performance and can be justified conceptually even when the indicators are not highly correlated with one another.

Objective: To compare 2 basic approaches for calculating a composite measure: an extension of the most widely-used approach, which weights individual indicators based on the number of people eligible for the indicator (referred to as denominator-based weights, DBWs), and a Bayesian hierarchical latent variable model (BLVM).

Methods: Using data for 15 QIs from 3275 hospitals in the Hospital Compare database, we calculated hospital ranks using several versions of DBWs and 2 BLVMs. Estimates in 1 BLVM were driven by differences in variances of the QIs (BLVM1) and estimates in the other by differences in the signal-to-noise ratios of the QIs (BLVM2).

Results: There was a high correlation in ranks among all of the DBW approaches and between those approaches and BLVM1. However, a high correlation does not necessarily mean that the same hospitals were ranked in the top or bottom quality deciles. In general, large hospitals were ranked in higher quality deciles by all of the approaches, though the effect was most apparent using BLVM2.

Conclusions: Both conceptually and practically, hospital-specific DBWs are a reasonable approach for calculating a composite measure. However, this approach fails to take into account differences in the reliability of estimates from hospitals of different sizes, a big advantage of the Bayesian models.

Key Words: quality performance, quality measurement, Bayesian inference

(*Med Care* 2008;46: 778–785)

From the *Center for Organization, Leadership and Management Research, VA Boston Healthcare System, Boston, Massachusetts; †School of Management, Boston University, Boston, Massachusetts; ‡International Business School, Brandeis University, Waltham, Massachusetts; and §Health Policy Institute, Boston University, Boston, Massachusetts.

This work was supported in part by a grant from the Commonwealth Fund.

Reprints: Michael Shwartz, PhD, School of Management, Boston University, 595 Commonwealth Avenue, Boston, MA 02215. E-mail: mshwartz@bu.edu.

ISSN: 0025-7079/08/4608-0778

Spurred by the Institute of Medicine,[1] the field of quality measurement has been growing rapidly. A number of consortia,[2] states,[3,4] and the federal government[5] publish reports on provider quality. One of the most prominent of these organizations is the Hospital Quality Alliance (HQA), a consortium of government agencies and private groups. As a result of their involvement in HQA, most US acute care hospitals submit data to the Centers for Medicare and Medicaid Services (CMS) on the level of adherence to a number of evidence-based process measures. These process measures are considered to be quality indicators (QIs). They are publicly available on the Hospital Compare website created by CMS and HQA.

Studies have shown the value of individual QIs, particularly when publicly reported, in stimulating quality improvement.[6–12] Individual indicators are often combined to derive an aggregate or composite measure of quality for a condition.[13–16] However, the value of aggregating individual indicators across conditions to develop an overall measure of hospital quality is less clear, particularly if correlations across conditions are relatively low.[14] In what follows, we justify a single composite measure 2 ways: (1) based on the reasonableness of such a measure from an organizational perspective; and (2) based on the literature about the nature of constructs derived from individual variables.

The vision of a high reliability organization[17–20] is influencing health care. An example of this is the Pursuing Perfection program, whose goal was to "push quality improvement in health care to a brand new level by creating models of excellence at a select number of provider organizations that would redesign all of their major care processes."[21] This level of improvement requires organizational transformation.[22,23] A hospital-wide measure of quality is a useful summary of progress along the road to this type of change. It reflects the extent to which top management has created a culture of quality and set of processes to ensure quality that have spread throughout the hospital. It helps to keep top managers focused on the "big picture" and offers a metric that can be used to monitor and compare performance across hospitals.

To provide an analogy, consider the measurement of financial performance. The single most important measure of an organization's financial performance is generally its profit margin. Although there is no doubt about the value of measuring profit margins for organizational subunits, there is equally no doubt about the value of aggregating subunits' margins into a composite organization-wide profit margin. From the perspective of top managers, who are responsible for overall organizational performance, attention first must be directed toward the organization's overall profit margin and then toward margins of subunits. Moreover, in studies of comparative organizational performance, an organization's overall profit margin is almost invariably the major measure used. Even if there were well done studies that showed little correlation in profit margins of organizational subunits (eg, clinical services) across a large number of similar organizations (eg, hospitals), it is unlikely that this would detract from interest in over-all profit margins. We believe that if quality is to assume the same level of importance as finances, the field needs bottom-line measures of quality performance comparable to bottom-line measures of financial performance.

The creation of a single composite measure of quality, irrespective of correlation among components of the measure, finds justification in the literature on the relation between constructs and measures. "In this literature, constructs are usually viewed as causes of measures, meaning that variation in a construct leads to variation in its measures. Such measures are termed reflective because they represent reflections, or manifestations, of a construct . . .. In some situations, measures are viewed as causes of constructs. Such measures are termed formative, meaning the construct is formed or induced by its measures. Formative measures are commonly used for constructs conceived as composites of specific component variables . . .."[24] As discussed by Dijkers et al,[25] Feinstein[26] brought this distinction to medicine, coining the term "clinimetrics." Clinicians often are interested in measuring a construct that combines multiple dimensions into a single score. The Apgar score, which combines 5 factors not correlated with each other into a single measure of newborn survivability, is a good example of a widely used formative score in medicine.

A single composite measure of quality would greatly facilitate implementation of pay-for-performance programs and value-based purchasing.[27,28] A number of different approaches for developing a composite measure of provider performance have been studied.[29–33] None, however, has been used to develop a single composite from the Hospital Compare QIs. In addition, a potentially valuable approach that has not been widely used but which has some advantages over existing approaches is a Bayesian hierarchical latent variable model (BLVM).[34] In this article, we describe 2 Bayesian models for estimating overall hospital performance from individual QIs, compare estimates of hospital performance from the models to estimates using the approach recommended by CMS for combining individual QIs into condition-specific measures, and discuss reasons for the differences in how hospitals are ranked.

METHODS

Data

We downloaded data from the CMS website[35] after the September 2006 update (which contained data for calendar year 2005) on QIs associated with acute myocardial infarction (AMI), congestive heart failure and pneumonia (see Appendix 1 for the indicators).

Current Approach for Developing Composite Measures

For each QI, the number of people eligible for the intervention reflected by the QI (denominator) and the number who receive the intervention (numerator) are known. The approach recommended by CMS for aggregating QIs within condition is to sum the numerators, sum the denominators, and then calculate the ratio of summed numerators to summed denominators.[36] This is equivalent to calculating a weighted average of the proportion eligible (ie, meet the medical criteria) for the QI who receive the intervention, where the weight applied to an indicator is the ratio of the number eligible for the indicator to the sum of the number eligible for all indicators (Appendix 2). We denote this approach, which uses denominator-based weights, as DBW. We consider several variations of this approach: (1) that recommended by CMS, which uses the number of cases at each hospital to derive hospital-specific denominator-based weights (DBWhs); (2) that recommended as one of the options by the Agency for Healthcare Research and Quality for deriving a composite measure from the Patient Safety Indicators, which sums the denominators across all the hospitals and calculates 1 set of denominator-based weights from this aggregation (DBWall)[37]; and (3) calculating separate DBW for large (400 or more beds), medium (100–399 beds), and small (25–99 beds) hospitals (DBWsize).

Shrinkage Estimators

Traditionally, methods for estimating a mean of an individual unit have been based on data from that particular unit. In an excellent nontechnical article, Efron and Morris[38] justify an alternative approach to estimation in which estimates about a unit are based on both data from that unit and on data from some larger set of units of which the particular unit is a member. For example, a provider's "true" performance on a particular QI might be estimated as a weighted average of the provider's observed performance on that indicator and the performance of all providers on that indicator (ie, the estimate of the provider's performance is "pulled" or "shrunken" toward the overall performance level based on data from all providers). The amount of shrinkage depends both on the reliability of observed performance by that provider, which to a large extent is a function of sample size, and on how far the provider's observed performance is from the performance level of all providers. A number of articles discuss and illustrate the value of these types of shrinkage estimators.[39–46] Hierarchical models generalize the idea of shrinkage and provide a comprehensive framework for examining clustered data. The nature of shrinkage occurring in these models is more complex than in the simple situation illustrated above, but the fundamental principle of basing estimates for an individual unit on data from that unit and from some wider set of units is the same. Bayesian hierarchical models are distinguished primarily by the way in which parameters are estimated.

Two Bayesian Hierarchical Latent Variable Models for Estimating a Composite Measure of Quality

The first model we consider is similar to Landrum et al.[34] We assume that there is an unobserved latent measure of quality at each hospital, which we denote by θ_h, where h indexes hospitals. We assume θ_h is normally distributed with mean 0 and variance 1 (ie, $\theta_h \sim N(0,1)$. Let t_{qh} be the unobserved "true" level of performance on QI q at hospital h. We assume:

$$\text{logit}(t_{qh})|a_q^0, a_q^1, \theta_h = a_q^0 + a_q^1\theta_h.$$

a_q^0 is a scaling factor that reflects differences in baseline values of the indicators. a_q^1 reflects the strength of the relationship between a specific indicator q and the latent measure of quality θ_h. To complete the model, let d_{qh} = the number of patients who receive indicator q at hospital h and n_{qh} be the number of eligible patients for the indicator. We assume $d_{qh}|t_{qh} \sim \text{binomial}(t_{qh}, n_{qh})$. This model differs from Landrum et al because, for analytical convenience, we use a logit function to link t_{qh} to θ_h rather than a probit function. This difference has no effect on estimates of hospital performance.

The above model assumes that all of the variation in t_{qh} is due to variation in θ_h. We also consider an extension of this model, specifically,

$$\text{logit}(t_{qh})|a_q^0, a_q^1, \theta_h, s_q^2 \sim N(a_q^0 + a_q^1\theta_h, s_q^2).$$

In this model, s_q reflects random variation in logit (t_{qh}) that is not due to its relationship to θ_h. We refer to the first Bayesian latent variable model as BLVM1 and the second as BLVM2.

The key driver in BLVM1 is a_q^1. To a large extent, differences in the a_q^1s reflect differences in variation of the QIs across hospitals. The key driver in BLVM2 is the ratio a_q^1/s_q. This is a signal-to-noise ratio that reflects the extent to which indicator q is correlated with the underlying latent variable.

To estimate model parameters, we used Gibbs sampling as implemented in WinBUGs.[47] This Markov Chain Monte Carlo estimation method generates samples of model parameters from the posterior distribution of the parameters, given the data and prior distributions of the parameters. We placed "flat" priors on the parameters, so the posterior distributions are driven by the data. We used as point estimates of the parameters the average of the values from the Gibbs samples.

Analysis

We fit the Bayesian models to 3275 hospitals on the Hospital Compare database that had 25 or more beds. However, in the results presented, we eliminated 64 hospitals where the sum of the denominators was less than 100. We examined the correlation of hospital ranks based on a composite measure calculated using the following approaches: the 3 denominator-based approaches (DBWhs, DBWall, DBWsize); 2 Bayesian approaches (BLVM1 and BLVM2); and 2 approaches that use weights derived from the Bayesian models. The Bayesian models do not estimate a set of weights comparable to the DBW. However, the model formulations imply a reasonable set of weights: for BLVM1, the a_q^1s, and for BLVM2, the a_q^1/s_q ratios. We refer to these as Bayesian-estimated weights (BEWs). To calculate a composite score from the weighted average of the observed adherence percentages, BEW1 uses the a_q^1s as weights and BEW2 uses the a_q^1/s_q ratios.

To examine the effect of each approach on estimates of quality for large, medium, and small hospitals, we analyzed the distribution of large, medium, and small hospitals across quality score deciles. We then examined the weights associated with each approach to understand why the different approaches resulted in different quality rankings for the large, medium, and small hospitals. Ranks based on the Bayesian models reflect both the model-implied weights and shrinkage. By comparing ranks from the Bayesian model to ranks when just using model-implied weights, we illustrate shrinkage in the Bayesian model estimates.

RESULTS

Table 1 shows the correlation of hospital ranks based on the different approaches. Most noticeable are the high correlations among ranks from the DBW approaches and the high correlation of these ranks with those from the 2 approaches based on Bayesian model 1 (ie, one just using the BEW1 and the other the full model (BLVM1)). Correlations are lower with the ranks from the 2 approaches based on Bayesian model 2 (ie, BEW2 and BLVM2).

A high correlation does not necessarily mean that the same hospitals are ranked in the extreme deciles (Table 2). The correlation between DBWhs ranks and BLVM1 ranks is 0.92. Nevertheless, only about 81% of hospitals ranked in the top decile by 1 method are ranked in the top decile by the other. However, most of the hospitals ranked differently are ranked only 1 decile lower. The correlation between DBWhs ranks and BLVM2 ranks is 0.77. As shown in Part B of Table 2, in this case only about 52% of hospitals ranked in the top decile by 1 method are ranked in the top decile by the other. Roughly half of those classified differently are classified more than 1 decile away.

Table 3 shows the distribution of large, medium, and small hospitals across quality deciles when different ap-

TABLE 1. Correlation of Ranks Calculated Using the Different Approaches

	Approach Used to Calculate Ranks						
	DBWhs	DBWall	DBWsize	BLVM1	BEW1	BLVM2	BEW2
DBWhs	1.00	0.94	0.95	0.92	0.89	0.77	0.78
DBWall	0.94	1.00	0.99	0.90	0.94	0.81	0.87
DBWsize	0.95	0.99	1.00	0.91	0.93	0.79	0.83
BLVM1	0.92	0.90	0.91	1.00	0.87	0.73	0.71
BEW1	0.89	0.94	0.93	0.87	1.00	0.73	0.82
BLVM2	0.77	0.81	0.79	0.73	0.73	1.00	0.89
BEW2	0.78	0.87	0.83	0.71	0.82	0.89	1.00

DBWhs indicates hospital-specific denominator-based weights; DBWall, all-hospital denominator-based weights, which are derived from an aggregation of cases across all hospitals; DBWsize, size-specific denominator-based weights, which are derived from an aggregation of cases of large, medium and small hospitals; BLVM1, Bayesian latent variable model 1; BEW1, Bayesian-estimated weights implied by Bayes latent variable model 1; BLVM2, Bayesian latent variable model 2; BEW2, Bayesian-estimated weights implied by Bayes latent variable model 2.

TABLE 2. Percent of Cases in Top and Bottom Deciles Using the Different Approaches to Calculate Ranks

Part A* DBWhs Deciles: Columns BLVM1 Deciles: Rows			Part B† DBWhs Deciles: Columns BLVM2 Deciles: Rows		
Top Deciles	**1**	**2**	**Top Deciles**	**1**	**2**
1	81.1	18.3	1	51.9	20.2
2	15.9	46.4	2	25.5	26.2
No. cases in decile	321	321		322	321
Bottom Deciles	**9**	**10**	**Bottom Deciles**	**9**	**10**
9	45.2	22.1	9	28.4	24.0
10	19.3	74.5	10	19.3	59.5
No. cases in decile	321	321		321	321

*Hospital-specific denominator-based weights (DBWhs) and Bayesian latent variable model 1 (BLMV1).

†Hospital-specific denominator-based weights (DBWhs) and Bayesian Latent Variable Model 2 (BLMV2).

proaches are used for ranking. The most consistent finding is that large hospitals tend to be ranked in higher deciles and small hospitals in lower deciles. Using size-specific DBWs rather than hospital-specific weights has little impact on the distributions. It is interesting, however, that the use of 1 set of DBWs (DBWall) helps the small hospitals somewhat at the expense of the large hospitals (9.6% of small hospitals are classified in the top decile versus 6.8% of large hospitals). The BLVM1 also helps the small hospitals at the expense of the large hospitals. BLVM2, however, does the opposite, helping the large hospitals at the expense of the small hospitals (16.2% of large hospitals are in the top decile versus only 4.9% of small hospitals).

Table 4, Part A shows the DBWall and DBWsize weights. At larger hospitals, a higher percentage of all cases are eligible for the AMI QIs (measures 1–6) and at smaller hospitals a larger percentage are eligible for the pneumonia QIs (measures 11–15). The sixth column shows the average percent adherence to each QI. In the last row of that column is the average of the individual QI adherence percents, which shows that across all hospitals and QIs, 78.9% of eligible cases received the QI intervention. The last row of columns 7 through 9 shows the overall average percent adherence by hospital size. The cells in these columns show the ratio of percent adherence to the QI to the average overall percent adherence. Thus, for example, the percent adherence to QI1 in large hospitals is $81.6 \times 1.17 = 95.5$. Of particular interest, small hospitals do relatively better (compared to the other size categories) on pneumonia QIs and large hospitals do relatively better on AMI QIs.

Table 4, part B shows BEW1 and BEW2 weights. In general, those measures with larger variances have larger BEW1 weights (as seen with measures 6 and 9, and to a lesser extent with measures 10, 12, and 15). However, as shown with measure 3 (as seen in Part A, DBWall, only 0.8% of all eligible cases are eligible for this measure), this relationship is mitigated somewhat by the prevalence of

TABLE 3. Percent of Cases in Each Quality Decile by Hospital Size and Approach Used to Calculate Ranks

Decile	DBWhs	DBWsize	DBWall	BLVM1	BLVM2
Large hospitals					
1	11.3	11.7	6.8	8.6	16.2
2	13.9	12.8	10.9	9.4	18.4
3	13.2	14.3	12.4	13.2	15.8
4	16.5	15.8	11.7	10.2	14.3
5	12.8	13.9	15.4	14.7	11.3
6	9.8	13.9	16.5	11.7	14.3
7	7.1	5.6	9.8	14.7	7.1
8	7.5	4.9	7.1	8.3	0.8
9	4.5	4.9	6.4	6.0	1.5
10	3.4	2.3	3.0	3.4	0.4
Medium hospitals					
1	11.2	10.9	10.7	10.4	12.0
2	10.6	11.1	11.4	10.8	11.7
3	10.7	10.4	10.8	10.3	11.9
4	10.4	10.5	11.4	11.7	11.1
5	10.5	10.4	10.4	10.1	11.7
6	10.3	10.3	9.8	9.5	10.5
7	10.2	10.3	10.6	10.0	9.0
8	9.8	9.6	9.2	10.2	9.5
9	9.0	9.6	9.4	9.8	7.2
10	7.3	6.9	6.4	7.1	5.3
Small hospitals					
1	7.7	8.1	9.6	9.7	4.9
2	7.9	7.3	7.3	8.7	4.8
3	7.9	8.2	7.9	8.7	5.0
4	7.6	7.6	7.1	6.8	6.9
5	8.5	8.3	8.0	8.5	6.6
6	9.5	8.5	8.7	10.5	8.0
7	10.4	10.5	9.0	8.7	12.5
8	10.9	12.0	12.2	10.1	13.3
9	13.2	12.1	12.1	11.3	17.1
10	16.5	17.5	18.1	16.8	20.8

the QI. Column 5 shows shared common variance of each QI (calculated as the R^2 from a model that predicts the QI percent adherence from all of the other QI percent adherences). In general, those measures with greater shared variance receive higher BEW2 weights (measures 1, 2, 4, and 5). Though measures 10 and 15 appear to have high shared variance, this results from the fact that these are the same QI (counseling in smoking cessation), but for different conditions. The important point to note is that BEW1 weights the pneumonia QIs relatively highly whereas BEW2 weights the AMI QIs relatively highly. Because large hospitals have a higher proportion of AMI cases and smaller hospitals a higher proportion of pneumonia cases, BEW2 and BLVM2 estimates of quality are relatively higher in the larger hospitals and BEW1 and BLVM1 estimates relatively higher in smaller hospitals.

In Table 5, we illustrate shrinkage. Hospitals are stratified into quintiles using the BEWs (a_1 for model 1 and a_1/s for model 2). The cells of the table in columns 2 and 3 show for hospitals in that decile the average of the following

TABLE 4. Denominator-Based Weights (A) and Bayesian-Estimated Weights (B)

Part A: Denominator-Based Weights and Adherence Percentages

	Denominator-Based Weights					Relative Adherence		
QI	DBWall	DBWlarge	DBWmed	DBWsmall	Average Adherence (%)	Large	Medium	Small
1	0.061	0.066	0.062	0.043	92.1	1.17	1.17	1.17
2	0.066	0.095	0.064	0.028	88.6	1.17	1.13	1.10
3	0.008	0.013	0.008	0.004	80.1	1.03	1.01	1.00
4	0.067	0.098	0.065	0.029	87.7	1.16	1.12	1.08
5	0.051	0.054	0.052	0.038	86.2	1.13	1.11	1.05
6	0.021	0.034	0.020	0.006	78.5	1.08	1.01	0.88
7	0.139	0.139	0.141	0.131	83.3	1.14	1.09	0.98
8	0.027	0.033	0.026	0.020	79.9	1.01	1.01	1.02
9	0.102	0.103	0.104	0.087	52.1	0.66	0.67	0.66
10	0.022	0.025	0.021	0.017	73.9	0.95	0.95	0.91
11	0.130	0.100	0.129	0.184	98.7	1.22	1.23	1.31
12	0.078	0.058	0.078	0.115	54.2	0.61	0.66	0.76
13	0.107	0.083	0.107	0.147	74.9	0.82	0.91	1.06
14	0.094	0.075	0.096	0.117	82.1	0.98	1.02	1.09
15	0.026	0.022	0.026	0.035	71.3	0.86	0.91	0.91
Overall average adherence (%)					78.9	81.6	80.5	75.2

Part B: Bayesian-Estimated Weights and Variances

	Bayesian-Estimated Weights*			
QI	Model 1	Model 2	Variance	Shared Variance†
1	0.040	0.100	0.011	0.482
2	0.064	0.126	0.023	0.568
3	0.041	0.095	0.055	0.227
4	0.066	0.194	0.026	0.633
5	0.049	0.126	0.024	0.580
6	0.110	0.057	0.075	0.386
7	0.068	0.070	0.029	0.397
8	0.039	0.058	0.030	0.316
9	0.135	0.034	0.071	0.423
10	0.109	0.037	0.055	0.651
11	0.069	0.027	0.001	0.144
12	0.080	0.023	0.059	0.380
13	0.023	0.007	0.016	0.225
14	0.020	0.017	0.008	0.135
15	0.086	0.028	0.053	0.605

*Rescaled so that the weights sum to 1.
†R^2 from a model predicting the QI percent adherence from the percent adherences of other quality indicators (technically, the communality from a factor analysis fit using maximum likelihood).

quantity: rank using BEWs minus rank from the full Bayesian model (BLVM). Because lower number ranks indicate higher quality (eg, rank "1" is the highest quality hospital), a negative number in the cell indicates that quality estimated by the BLVM is lower than quality measured by the BEWs (which does not include shrinkage); a positive number in the cell indicates quality from the BLVM is higher than quality using the BEWs. Shrinkage is apparent. In the top 2 deciles, on average quality estimates from the BLVMs are lower than estimates from the BEWs; in the bottom 2 deciles, on average quality estimates from the BLVMs are higher than from the BEWs. Thus, on average, shrinkage (which occurs when using the BLVM but not the BEWs) pulls estimates of quality in the off-center deciles toward the middle. In the fourth column, we consider the absolute value of the change in ranks (so that positive and negative changes do not cancel each other out) and its relationship to the total number of cases at the hospital. There is a negative relationship. When a hospital has more cases, there are fewer rank changes when BLVM estimates are compared with BEW estimates.

DISCUSSION

For most of the approaches we examined, hospital ranks were highly correlated. Hence, in terms of assessing organiza-

TABLE 5. Illustrating Shrinkage by Comparing Ranks Based on Bayesian-Estimated Weights to Ranks from the Bayesian Models

Bayesian-Estimated Weight (BEW) Quintile	Average of BEW2 Rank Minus BLVM2 Rank	Average of BEW1 Rank Minus BLVM1 Rank	Correlation of Number of Eligible Cases in Hospital and Absolute Value of BEW1 Rank Minus BLVM1 Rank
1	−98.1	−79.7	−0.12*
2	−53.9	−86.2	−0.26*
3	35.2	−57.6	−0.21*
4	66.9	76.4	−0.30*
5	50.5	147.6	−0.09†

*Significant at $P = 0.01$ level.
†Significant at $P = 0.05$ level.

tional performance and monitoring changes in performance over time, the choice of approach is unlikely to make much difference. The only time when it does matter is in the context of pay-for-performance programs that set a specific threshold affecting payments or when emphasis is placed on those above or below specific thresholds, such as in the case of public reports. As we have shown, even though ranks are highly correlated, different approaches result in differences in the specific hospitals that are in the top or bottom deciles.

An underlying assumption of the approaches considered is that all of the QIs are of equal clinical importance. The incentives created by hospital-specific DBW are most consistent with this assumption. As described above, using DBWhs is equivalent to calculating the composite measure as the ratio of the sum of all the QI numerators to the sum of all the QI denominators. Thus, an increase of 1 case in the numerator of any of the QIs has exactly the same effect on the composite measure. None of the other approaches has this desirable quality. DBW derived from the aggregation of either all hospitals or sets of hospitals (eg, based on size) are useful because they allow one to compare weights across the sets of hospitals or to weights from other approaches. However, aside from this advantage, it is not clear why one would use DBW from aggregations of hospitals.

As discussed in the Introduction, if the composite measure is considered a formative scale, justification for creating the composite does not depend on correlation among the individual components of the scale. The Bayesian models imply a reflective scale. a^1 is the parameter that links each observed QI to the latent variable. Although we did not emphasize this in the Results section, none of the intervals within which the a^1s fell with 95% certainty overlapped zero. This indicates that all of the QIs had a statistically significant relationship to the underlying latent variable. Hence, at least statistically, there is sufficient correlation to justify an underlying latent trait.

In BLVM1, once we "know" a hospital's latent variable (ie, θ_h), its "true" level of performance on QI q (ie, t_{qh}) is determined by the relationship $a_q^0 + a_q^1\theta_h$. BLVM2 hypothesizes that there is some random variation associated with each measure q, measured by s_q, in addition to variation induced by the QI's link to θ_h. As discussed, the different formulations result in different implied weights. In BLVM1, the implied weights are, to a large extent, related to differences in the variance of the observed QIs; in BLVM2, they are related to the extent to which observed data reflect a "signal" about the underlying trait versus noise (ie, to the ratio a^1/s). On the one hand, it seems reasonable to give more weight to QIs with greater variation because that is where there is more opportunity to learn about differences in quality. On the other hand, it is possible that differences in variation in the QIs may reflect differences in consensus about the clinical importance of the measures. Although the assumption of equal clinical importance underlies all of these approaches, in practice the possibility that there is not similar consensus might cause second thoughts about basing weights on differences in variance.

Although weights based on differences in the signal-to-noise ratio hold some conceptual appeal, the practical implications of this choice are not that attractive, at least when using the QIs we considered. In general, all of the approaches indicate that the larger hospitals have higher quality and the smaller hospitals lower quality. However, as shown in Table 3, BLVM2 results in the largest differences in quality estimates between large and small hospitals, mainly due to the higher weights given to the AMI QIs at the expense of the pneumonia QIs.

One might ask the question: "Why use the Bayesian models at all?" There are several answers to this. The first, which we have emphasized in this article, is that the shrinkage estimates from the Bayesian models appropriately take into account differences in the reliability of estimates from hospitals of different sizes. These differences are not reflected in composite measures resulting from DBW. The second, which we have not emphasized here is that probability intervals can be placed around performance measures like ranks, which usually highlight that there is a great deal of uncertainty associated with ranks.[48,49] Pay-for-performance programs thus far have not formally taken into account the issue of uncertainty in estimates of performance. Finally, Bayesian models allow estimates of policy-relevant performance in ways that other approaches do not. For example, one can calculate the probability that each hospital exceeds or falls below thresholds of interest.

In its current options article on Medicare hospital value-based purchasing, CMS proposes that for each QI a hospital is assigned points based both on the adherence percentage and on improvements in the percentage.[50] A composite is calculated by summing points across QIs and dividing by the total number of points possible. It is worth noting that this approach does not take into account: (1) differences in the importance of the indicator at a specific hospital; (2) the effect of sample sizes on the reliability of adherence percents for the indicator; and (3) uncertainty associated with the resulting summary statistics. It would be useful to consider Bayesian models as an approach for calculating a total value-based purchasing performance score from the individual QI scores.

Concerns have been raised about the extent to which the Hospital Compare QIs reflect differences in documentation rather than differences in quality.[51,52] In addition, the particular set of QIs we considered is limited to 3 conditions, 2 of which are heart related. As noted, this affects our conclusions. However, our purpose is not to argue for the validity of rankings based on these particular indicators, but rather to illustrate some of the issues involved when using Bayesian latent variable models as compared with DBW to calculate a composite measure of quality. For this purpose, concerns about the QIs are less important.

ACKNOWLEDGMENTS

The authors thank the 3 anonymous reviewers for their thoughtful questions, comments and suggestions that made a large contribution to the quality of the final manuscript.

REFERENCES

1. Institute of Medicine, Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington, DC: National Academy Press; 2001.
2. Leapfrog Group. *The Leapfrog Group Hospital Patient Safety Survey, April 2003–March 2004*. Washington, DC: Leapfrog Group; 2004.
3. Pennsylvania Health Care Cost Containment Council. *Hospital Performance Report*. Harrisburg, PA: Pennsylvania Health Care Cost Containment Council; 2005.
4. California CABG Mortality Reporting Program. The California Report on Coronary Artery Bypass Graft Surgery: 2000–2002 Hospital Data. Available at: http://www.oshpd.cahwnet.gov/HQAD/Outcomes/Studies/cabg/2000-2002Report/index.htm.
5. Centers for Medicare & Medicaid Services. Nursing Home Compare. Available at: http://www.medicare.gov/NHCompare and Hospital Compare: http://www.hospitalcompare.hhs.gov/.
6. Marciniak TA, Ellerbeck EF, Radford MJ, et al. Improving the quality of care for Medicare patients with acute myocardial infarction: results from the Cooperative Cardiovascular Project. *JAMA*. 1998;279:1351–1357.
7. Marshall MN, Shekelle PG, Leatherman S, et al. The public release of performance data: what do we expect to gain? A review of the evidence. *JAMA*. 2000;283:1866–1874.
8. Ferguson TB, Peterson ED, Coombs LP, et al. Use of continuous quality improvement to increase use of process measures in patient undergoing coronary artery bypass graft survery: a randomized controlled trial. *JAMA*. 2003;290:49–56.
9. Bradley EH, Holmboe ES, Mattera JA, et al. Data feedback efforts in quality improvement: lessons learned from US hospitals. *Qual Saf Health Care*. 2004;13:26–31.
10. Gibberd R. , Hancock S, Howley P, et al. Using indicators to quantify the potential to improve the quality of health care. *Int J Qual Health Care*. 2004;16(Suppl 1):i37–i43.
11. Williams SC, Schmaltz SP, Morton MS, et al. Quality of care in U.S. hospitals as reflected by standardized measures. *N Engl J Med*. 2005; 353:255–264.
12. Hibbard JH, Stockard J, Tusler M. Hospital performance reports: impact on quality, market share, and reputation. *Health Aff*. 2005;24:1150–1160.
13. Lindenauer PK, Remus D, Roman S, et al. Public reporting and pay for performance in hospital quality improvement. *N Engl J Med*. 2007;356:486–496.
14. Jha AK, Zhonghe L, Orav EJ, et al. Care in U.S. hospitals—The Hospital Quality Alliance Program. *N Engl J Med*. 2005;353:265–274.
15. Werner RM, Bradlow ET. Relationship between Medicare's hospital compare performance measures and mortality rates. *JAMA*. 2006;296:2694–2702.
16. Jha AK, Orav EJ, Zhonghe L, et al. The inverse relationship between mortality rates and performance in The Hospital Quality Alliance Measures. *Health Aff*. 2007;26:1104–1110.
17. Reason J. *Human Error.* Cambridge: Cambridge University Press; 1990.
18. Perrow C. *Normal Accidents: Living With High-Risk Technologies.* 2nd ed. Princeton: Princeton University Press; 1999.
19. Weick KE, Sutcliffe KM, Obstfeld D. Organizing for high reliability: processes of collective mindfulness. *Org Behav*. 1999;21:81–123.
20. Helmreich RL. Human error: models and management. *Br Med J*. 2000;320:768–770.
21. Berwick D, Kabcenell A, Nolan T. Pursuing perfection: no Toyota yet, but a start. *Mod Hlthc*. 2005;35:18–19.
22. NHS Institute for Innovation and Improvement, Matrix Research and Consultancy. *What is transformation change? Literature review.* Coventry, UK: University of Warwick; 2006.
23. Lukas CV, Holmes SK, Cohen AB, et al. Transformational change in health care systems: an organizational model. *Health Care Manage Rev*. 2007;32:309–320.
24. Edwards JR, Bagozzi RP. On the nature and direction of relationships between constructs and measures. *Psychol Meth*. 2000;5:155–174. (quote 155–156).
25. Dijkers MPJM, Diamond JJ, Marion R. Psychometrics and clinimetrics in assessing environments: a comment suggested by MacKenzie et al., 2002/ Replies. *J Allied Health.* 2003;32:38–45.
26. Feinstein AR. Multi-item "instruments" vs Virginia Apgar's principles of clinimetrics. *Arch Intern Med*. 1999;159:125–128.
27. Rowe JW. Pay-for-performance and accountability: related themes in improving health care. *Ann Intern Med*. 2006;145:695–699.
28. Rosenthal MB, Dudley RA. Pay-for-performance: will the latest payment trend improve care? *JAMA*. 2007;297:740–744.
29. Zaslavsky AM, Shaul JA, Zaborski LB, et al. Combining health plan performance indicators into simpler composite measures. *Hlth Care Fin Rev*. 2002;23:101–115.
30. Lied TR, Malsbary R, Eisenberg C, et al. Combining HEDIS indicators: a new approach to measuring plan performance. *Hlth Care Fin Rev*. 2002;23:117–129.
31. Jacobs R, Goddard M, Smith PC. How robust are hospital ranks based on composite performance measures. *Med Care*. 2005;43: 1177–1184.
32. Reeves D, Campbell SM, Adams J, et al. Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. *Med Care*. 2007;45:489–496.
33. Caldis T. Composite health plan quality scales. *Hlth Care Fin Rev*. 2007;28:95–107.
34. Landrum MB, Bronskill SE, Normand S-L. Analytic methods for constructing cross-sectional profiles of health care providers. *Heal Serv Out Res Meth*. 2000;1:23–47.
35. Available at: http://www.hospitalcompare.hhs.gov/Hospital/.
36. US Department of Health & Human Services. Information for professionals on data collection. Available at: http://www.hospitalcompare.hhs.gov/Hospital/Static/Data-Professionals.asp?dest=NAVHomeDataDetailsProfessionalInfo#TabTop.
37. Agency for Healthcare Research and Quality (AHRQ) Quality Indicators: Patient Safety Indicator Composite Measure Final Technical Report. October 2006.
38. Efron B, Morris C. Stein's paradox in statistics. *Scien Amer*. 1977;236: 119–127.
39. Shwartz M, Ash AS, Anderson J, et al. Small area variations in hospitalization rates: how much you see depends on how you look. *Med Care*. 1994;32:189–201.
40. Greenland S. Principles of multilevel modeling. *Int J Epidem*. 2000;29: 158–167.
41. Shahian DM, Normand S-L, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg*. 2001;72:2155–2168.
42. Arling G, Lewis T, Kane RL, et al. Improving quality assessment through multilevel modeling: the case of nursing home compare. *Health Serv Res*. 2007;42:1177–1199.
43. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med*. 1997;127:764–768.
44. Normand S-L, Glickman M, Gastsonis C. Statistical methods for profiling providers of medical care: issues and applications. *J Amer Stat Assoc*. 1997;92:803–814.

45. Burgess JF, Christiansen CL, Michalak SE, et al. Medical profiling: improving standards and risk adjustments using hierarchical models. *J Health Econ*. 2000;19:291–309.
46. Landrum MB, Normand S-L, Rosenheck RA. Selection of related multivariate means: monitoring psychiatric care in the Department of Veterans Affairs. *J Amer Stat Assoc*. 2003;98:7–16.
47. Spiegelhalter DJ, Thomas A, Best NG, et al. WinBUGS, version 1.4.1. Cambridge: MRC Biostatistics Unit; 2003.
48. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J Roy Stat Soc*. 1996;159:385–443.
49. Davidson G, Moscovice I, Remus D. Hospital size, uncertainty, and pay-for-performance. *Hlth Care Fin Rev*. 2007;29:45–57.
50. American Hospital Association Quality Advisory: CMS issues options paper on Medicare hospital value-based purchasing. April 12, 2007.
51. Laschober M, Maxfield M, Felt-Lisk S, et al. Hospital response to public reporting of quality indicators. *Hlth Care Fin Rev*. 2007;28:61–76.
52. Goldman LE, Henderson S, Dohan DP, et al. Public reporting and pay-for-performance: safety-net hospital executives' concerns and policy suggestions. *Inquiry*. 2007;44:137–145.

APPENDIX 1

Evidence-Based Indicators for Acute Myocardial Infarction, Congestive Heart Failure and Pneumonia (in the text, we refer to the indicators by the numbers to the left of each indicator)

Acute Myocardial Infarction (AMI)

1. Aspirin at arrival
2. Aspirin prescribed at discharge
3. Angiotensin-converting enzyme (ACE) inhibitor or angiotensin receptor blocker (ARB) for left ventricular systolic dysfunction (LVSD)
4. Beta blocker prescribed at discharge
5. Beta blocker at arrival
6. Adult smoking cessation advice/counsel

Congestive Heart Failure

1. Left ventricular function assessment
2. ACE inhibitor or ARB for LVSD
3. Discharge instructions
4. Adult smoking cessation advice/counseling

Pneumonia

1. Oxygenation assessment
2. Pneumococcal vaccination status assessment
3. Initial antibiotic received within 4 hours of hospital arrival
4. Blood culture performed in emergency department before first antibiotic received in hospital
5. Adult smoking cessation advice/counseling

APPENDIX 2

Denominator-Based Weights

Let d_{qh} = the number of patients who receive QI q at hospital h and n_{qh} = the number of patients eligible for QI q at hospital h. The proportion of eligible patients who receive QI q at hospital h is (d_{qh}/n_{qh}). The DBW for QI q at hospital h is $(n_{qh}/\Sigma_q n_{qh})$. The composite measure of quality at hospital h using DBW is:

$$\Sigma_q[(n_{qh}/\Sigma_q n_{qh}) \times (d_{qh}/n_{qh})] = \Sigma_q[d_{qh}/\Sigma_q n_{qh}] = \Sigma_q d_{qh}/\Sigma_q n_{qh}.$$