A Probability Metric for Identifying High-Performing Facilities

An Application for Pay-for-Performance Programs

Michael Shwartz, PhD,† Erol A. Peköz, PhD,*† James F. Burgess, Jr, PhD,*‡ Cindy L. Christiansen, PhD,‡ Amy K. Rosen, PhD,*§ and Dan Berlowitz, MD, MPH‖‡*

Background: Two approaches are commonly used for identifying high-performing facilities on a performance measure: one, that the facility is in a top quantile (eg, quintile or quartile); and two, that a confidence interval is below (or above) the average of the measure for all facilities. This type of yes/no designation often does not do well in distinguishing high-performing from average-performing facilities.

Objective: To illustrate an alternative continuous-valued metric for profiling facilities—the probability a facility is in a top quantile—and show the implications of using this metric for profiling and pay-for-performance.

Methods: We created a composite measure of quality from fiscal year 2007 data based on 28 quality indicators from 112 Veterans Health Administration nursing homes. A Bayesian hierarchical multivariate normal-binomial model was used to estimate shrunken rates of the 28 quality indicators, which were combined into a composite measure using opportunity-based weights. Rates were estimated using Markov Chain Monte Carlo methods as implemented in WinBUGS. The probability metric was calculated from the simulation replications.

Results: Our probability metric allowed better discrimination of high performers than the point or interval estimate of the composite score. In a pay-for-performance program, a smaller top quantile (eg, a quintile) resulted in more resources being allocated to the highest performers, whereas a larger top quantile (eg, being above the median) distinguished less among high performers and allocated more resources to average performers.

Conclusion: The probability metric has potential but needs to be evaluated by stakeholders in different types of delivery systems.

Key Words: profiling, performance measurement, quality measurement, Bayesian models, pay-for-performance

(*Med Care* 2014;52: 1030–1036)

There are 2 commonly used approaches for identifying high-performing facilities on measures such as 30-day mortality and readmission rates or adherence to processes of care. The first considers whether the facility is in a top quantile (eg, quintile, quartile, or above the median)[1–3]; and the second determines if the confidence interval associated with the measure is below (or above if high values of the measure are good) the average of the measure for all facilities.[4] Both approaches flag some facilities as high performing and others as not. However, there are problems with this type of yes/no classification. Assume, of 100 facilities we identify the top 20 as high performing and, in the context of a pay-for-performance (P4P) program, give them a bonus or use them as benchmarks for lower-performing facilities. Do we really believe that the 19th and 20th ranked facilities are significantly better than the 21st and 22nd ranked facilities? In fact, there almost always is a much larger gap in performance between the top-ranked facility in a quantile and the bottom-ranked facility in the quantile than between the bottom-ranked facility and the top-ranked facilities in the next quantile. Yet, the yes/no approach to identifying high performers ignores the former gap and highlights the latter. The same problem arises when facilities are identified as high performing because their confidence interval is below (or above) the mean of all facilities. Although we can identify a set of facilities that we are fairly confident is better than average, the confidence intervals associated with these high-performing facilities often have substantial overlap with the confidence intervals of average-performing facilities (ie, those that include the mean). Thus, we are not able to conclude that many of the high-performing facilities are measurably different from the average-performing facilities.

In this paper, we illustrate an alternative continuous-valued metric for profiling facilities—the probability that a

From the *Center for Healthcare Organization and Implementation Research, VA Boston Healthcare System (152M); †School of Management; ‡School of Public Health; §Department of Surgery, School of Medicine, Boston University; and ‖Center for Healthcare Organization and Implementation Research, Bedford VA Hospital, Bedford, MA.

Supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development IIR 06-260. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

The authors declare no conflict of interest.

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

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Website, www.lww-medicalcare.com.

ISSN: 0025-7079/14/5212-1030

facility is in a top quantile—and show implications of using this metric for profiling and P4P programs. The probability measure allows us to rank facilities and thus identify those in a top quantile. However, it also indicates, using the same single number, how confident we should be about its high-performer designation. In a P4P program, if funds are allocated to facilities in proportion to the probability a facility is a high performer, all facilities may receive something, but those facilities with a higher likelihood of being a high performer will receive more. This is a conceptually more appealing way of allocating resources than an approach that designates certain facilities as high performers and allocates all of the resources to them.

The Centers for Medicare and Medicaid Services (CMS) P4P program, called a value-based purchasing program, was created to incentivize high quality hospital care. It does not use a designation of high-performing status to dichotomize hospitals, but maps a hospital's performance on each measure into a measure-specific score, which is then aggregated across measures into an overall score that determines payment adjustments.[5] The problem with the CMS approach is that it does not reflect the reliability of each score. For example, 85% adherence to a process of care measure is counted the same whether it is based on 30 eligible patients or 300. In the former case, the probability the facility in a top quantile is lower than in the latter case, reflecting the increased uncertainty associated with a measure based on only 30 versus 300 patients.

In its Medicare Advantage Program and on Nursing Home Compare,[6] CMS awards each facility from 1 to 5 stars based on the quintile in which an aggregation of their performance scores fall. The disadvantage of this approach is that consumers have no way of knowing the level of confidence associated with a star designation. Rather than focusing on the probability a facility is in some top quantile, one can calculate the probability a facility is in each of the quintiles, thus communicating the confidence in each star designation. In the Results section, we illustrate a possible way to portray this information.

Calculation of probability metrics is relatively straightforward when Markov Chain Monte Carlo (MCMC) methods are used to estimate model parameters. Several studies have illustrated probability metrics in a health care setting, including Normand et al,[7] who consider a continuous-valued metric and Christiansen and Morris[8] and Burgess et al,[9] who use the metric as a way to identify high-performing or low-performing facilities. However, the metric has not gained traction, which provides motivation for this paper. Our contribution is to compare the implications of using a probability metric to identify high-performing facilities to more commonly used approaches based upon ranks or confidence intervals; and to examine the implications of the metric in P4P programs. One likely reason for limited use of this approach is the complexity of MCMC methods; but, for those with only moderate statistical training, MCMC methods are no less complex or transparent than the hierarchical generalized linear models used in current CMS profiling.[4] In Supplemental Digital Content 1 (http://links.lww.com/MLR/A807, text), we give a brief nontechnical description of MCMC methods using a dice example to increase the transparency in the way in which the probabilities are generated.

METHODS

Setting

To illustrate our approach, we examined quality indicator (QI) data from 112 of the 132 Veterans Health Administration (VA) nursing homes (called Community Living Centers) that had at least 10 long-stay residents (over 90 d) and where at least one third of the residents were long stay (based on average daily census). We used data from fiscal year 2007 (October, 2006 through September, 2007).

QIs

We considered a set of 24 QIs calculated from the Minimum Data Set (MDS), version 2.0. The MDS is a core component of the Resident Assessment Instrument, an instrument whose use for quarterly patient assessment and care planning is required in all nursing homes that receive federal reimbursement. The Resident Assessment Instrument is also used in the VA. MDS data provide a summary assessment of the status of each long-stay nursing home resident.

Before the recent implementation of MDS 3.0, the 24 QIs were routinely provided to non-VA nursing homes and used by regulators as a preliminary step in the certification process.[10] Four of the indicators are stratified into high-risk and low-risk residents, resulting in a total of 28 QIs. Before implementation by the VA of MDS 3.0 in 2012, these indicators were provided monthly to VA Community Living Centers. Supplemental Digital Content 2 (http://links.lww.com/MLR/A808, Table) shows the 28 QIs, all of which are of the form "the proportion of residents eligible for the indicator (the denominator) who experience the event of interest (the numerator)." As the indicators measure adverse events (ie, events considered as related to care received in the facility), lower levels of the indicators indicate higher performance.

Analysis

To illustrate the probability metric, we used a composite measure that was calculated as a weighted average of the 28 individual QIs using opportunity-based weights (ie, the weight associated with each QI is the denominator for that QI divided by the sum of the denominators for all of the QIs). This approach, similar to that used by CMS in its P4P demonstration programs,[11,12] is equivalent to treating each of the individual QIs as equally important. Calculating a composite as a weighted average implies that the composite is conceptualized as a formative construct.[13] In a formative construct, one would not necessarily expect the different indicators to be correlated as they represent different dimensions of quality. Other than the high-risk/low-risk QIs, most of the other 28 QIs were not highly correlated with each other.

The denominators used to calculate the QIs varied across facilities (from 10 to 225 residents, median = 62) and within facility, across the QIs. Hence, there were substantial differences in the reliability of the observed rate in different facilities. A number of studies have shown the value of stabilizing rates in this type of situation by "shrinking" them toward

population averages,[14–20] including several particularly accessible to less technical readers.[15–18] Similar to Shwartz et al,[20] we used MCMC methods to estimate from a Bayesian hierarchical multivariate normal-binomial model the shrunken rates for the 28 QIs, which were then combined into a composite measure (Supplemental Digital Content 3, http://links.lww.com/MLR/A809, text). In Shwartz et al,[20] there is a simple illustration of how shrinkage works in this type of model. We also estimated 95% credible intervals for the composite measure. These intervals are the range within which we are 95% sure the composite measure lies. It is straightforward to calculate the probability metric from the Gibbs samples by ranking the composite measures in each sample and then assigning a 1 if a facility's rank is in the top quantile and a 0 otherwise. The probability a facility is in the top quantile is the proportion of 1's in the set of Gibbs samples.

We compared facility ranks and identification of high performers based on point estimates of the composite measure and their associated credible intervals to ranks and high-performer identification based on the probability metric. When using the probability metric, 2 cutoffs were used to define high performance: being in the top quintile and being above the median. We denote the probability of being in the top quintile by PTQ and the probability of being in the top half of all facilities by PTH.

We then examined the implications of the probability metric for a P4P program. When CMS profiles facilities in Hospital Compare, it identifies facilities as high or low performers based on whether or not their confidence interval includes the population average. However, in their value-based purchasing program, they reward facilities by converting a performance measure into a score and then aggregating the score across measures, which is then translated into a payment bonus.[5] We avoid the aggregation problem by considering a composite score. To allocate bonus payments in a simulated P4P program, we used an approach similar in spirit to that used by CMS. Specifically, to convert a facility's PTQ (or PTH) into a percentage of the total amount set aside for payment bonuses that each facility would receive, we took the ratio of the facility's PTQ (or PTH) to the sum of the PTQs (or PTHs) across all facilities.[21]

Finally, we conducted several sensitivity analyses, examining the impact of 2 different priors on estimates; the impact if the composite measures have a skewed distribution; and the impact if P4P is based upon the individual QIs rather than a composite measure. The details of the sensitivity analyses are described in Supplemental Digital Content 4 (http://links.lww.com/MLR/A810, text and Figures).

RESULTS

Figure 1 shows the interval estimates of facility performance based on the composite score from the 28 QIs, ordered from the highest performer (lowest score) to the lowest performer (highest score). There were 28 facilities that would be classified as high performers because their interval estimate was below the mean composite score of 0.128. As shown in the figure, for many of these facilities, particularly those with an interval estimate whose upper value was close to 0.128, their interval estimate overlapped noticeably with the interval estimates of other facilities classified as average performers (because their interval estimate included 0.128).

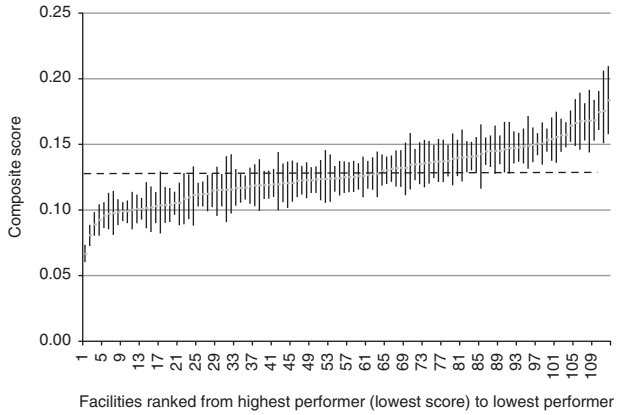

FIGURE 1. Interval estimates of performance ranked from highest-performing facility to lowest. It shows the 95% credible intervals (CIs) for facilities ranked from highest performance to lowest based on the point estimate of their composite score. The dashed line at 0.128 is mean performance across all facilities. It is apparent from the this that the 95% CIs of a number of high-performing facilities (those whose 95% CIs are <0.128) overlap with the 95% CIs of average-performing facilities (those whose 95% credible intervals cover 0.128). For example, the facility ranked 27th is a high performer (95% CI, 0.103–0.121), whereas the facility ranked 30th is an average performer (95% CI, 0.100–0.135). However, because of the overlap in their 95% CIs, there is no basis for saying the 27th ranked facility is better than the 30th ranked facility.

The performance statistics for the top 30 facilities ranked by PTQ (ie, the probability of being among the top 23) is shown in Table 1. The top 23 ranked facilities are the same as the top 23 facilities based on the point estimate of the composite score, although the order is slightly different. However, PTQ begins to decline as one moves toward lower-ranked facilities in the quintile. The top 10 facilities have over a 95% chance of being in the top quintile. However, the 17th ranked facility has only an 80% chance of being in the top quintile and 20th ranked facility only a 73% chance. The last 3 facilities in the quintile have under a 50% chance of being in the quintile. Thus, if we labeled the top 23 facilities as being high performers, it is more likely than not that 3 of the high-performing facilities were really average performers. Facilities ranked 24–30 (just below the quintile) have a very similar composite score to the facilities ranked 22nd and 23rd, but very different probabilities of being in the top quintile. Further, some of the high-performing facilities based on their interval estimates have a relatively low probability of being in the top quintile, for example, the 23rd, and 26th to 29th ranked facilities in Table 1. It seems difficult to justify treating all of these facilities the same based on the fact that their interval estimate is <0.128.

The probability a facility is in the top quintile, PTQ (smooth plot), ranked from the highest probability to the lowest, and the probability a facility is above the median,

TABLE 1. Performance Statistics for Top 30 Facililties Ranked by Probability of Being in the Top Quintile (PTQ)*

	Composite	Interval Estimate		Probability In	Probability In	Top Quintile	Top Half
Facility	Score	Lower	Upper	Top Quintile	Top Half	% of Pool	% of Pool
1	0.067	*0.061*	*0.073*	1.000	1.000	0.043	0.018
2	0.080	*0.073*	*0.088*	1.000	1.000	0.043	0.018
3	0.088	*0.079*	*0.097*	1.000	1.000	0.043	0.018
4	0.098	*0.090*	*0.106*	0.996	1.000	0.043	0.018
5	0.094	*0.084*	*0.105*	0.993	1.000	0.043	0.018
6	0.097	*0.086*	*0.107*	0.992	1.000	0.043	0.018
7	0.099	*0.091*	*0.107*	0.987	1.000	0.043	0.018
8	0.098	*0.087*	*0.108*	0.982	1.000	0.043	0.018
9	0.099	*0.090*	*0.110*	0.970	1.000	0.042	0.018
10	0.098	*0.085*	*0.112*	0.954	1.000	0.041	0.018
11	0.101	*0.091*	*0.112*	0.937	1.000	0.041	0.018
12	0.101	*0.087*	*0.114*	0.885	1.000	0.038	0.018
13	0.099	*0.079*	*0.119*	0.870	1.000	0.038	0.018
14	0.103	*0.092*	*0.115*	0.856	1.000	0.037	0.018
15	0.102	*0.087*	*0.118*	0.824	1.000	0.036	0.018
16	0.102	*0.088*	*0.116*	0.821	1.000	0.036	0.018
17	0.103	*0.091*	*0.116*	0.800	1.000	0.035	0.018
18	0.106	*0.098*	*0.114*	0.774	1.000	0.034	0.018
19	0.104	*0.089*	*0.122*	0.733	0.997	0.032	0.018
20	0.104	*0.087*	*0.122*	0.729	0.996	0.032	0.018
21	0.109	0.090	0.131	0.495	0.946	0.022	0.017
22	0.111	0.094	0.130	0.465	0.934	0.020	0.017
23	0.110	*0.094*	*0.127*	0.460	0.970	0.020	0.017
24	0.112	0.092	0.142	0.448	0.914	0.019	0.016
25	0.114	0.090	0.140	0.338	0.858	0.015	0.015
26	0.112	*0.101*	*0.125*	0.323	0.990	0.014	0.018
27	0.112	*0.103*	*0.121*	0.284	0.996	0.012	0.018
28	0.113	*0.102*	*0.125*	0.282	0.981	0.012	0.018
29	0.113	*0.103*	*0.123*	0.233	0.996	0.010	0.018
30	0.118	0.100	0.135	0.160	0.835	0.007	0.015

*For each facility the point estimate of the composite score from the Bayesian multivariate normal-binomial model, the 95% credible interval (interval estimate), the probability the facility is in the top quintile (PTQ) and top half (PTH) of all facilities, and the proportion of a P4P payment pool each facility would receive if payment were based on PTQ and PTH are shown. The intervals in italics are below the overall average PTQ (0.128) and would be flagged as high performers.

PTH (jagged plot) are shown in Figure 2A. All 23 highest ranked facilities have a $\geq 93\%$ chance of being above the median and 20 of the 23 have over a 99% chance, whereas the probability of being in the top quintile varies from 1.00 to 0.46. As seen in the figure, PTQ discriminates much more among high-ranked facilities than PTH. However, PTH discriminates much more among those facilities in the lower half of the performance distribution. The sharp drops in PTH occur for smaller facilities.

Figure 2B (and the last 2 columns of Table 1) translate the probabilities in Figure 2A into a P4P context. The greater discrimination among high-ranked facilities when using PTQ is reflected in the greater percentage of the payment allocation that facilities ranked in the top 11 would receive, above 4%, versus those ranked at the bottom of the quintile, around 2%. Using PTH, the top 23 ranked facilities would receive between 1.7% and 1.8% of the payment allocation. Using PTQ, payment declines rapidly for facilities ranked 21st to 49th, at which point facilities receive close to nothing (under 0.1%). Using PTH, the mid-ranked facilities (25th to 66th) would receive a noticeably higher payment than using PTQ and some of the bottom-ranked facilities would do somewhat better.

Table 2 illustrates extension of the probability metric from a focus on the top quintile to consideration of the probability of being in each of the quintiles. It shows for 20 "boundary" facilities (those facilities with ranks close to the ranks that define quintile 3) the probability that they are in each of the quintiles. Here we distinguish the quintiles by stars, where 5 stars indicate facilities in quintile 1—the highest ranked, and 1 star indicates those in quintile 5—the lowest ranked. The facilities ranked in the bottom 5 in quintile 2 (4 stars) based on the point estimate of their composite score have over a 45% of being 3-star or lower facilities, and 4 of the 5 facilities ranked in the bottom 4 in quintile 3 (3 stars) have over a 34% of being 2-star facilities. Thus, for these types of boundary facilities, there is a fair bit of uncertainty associated with their star designation. Figure 3, which shows bar charts indicating the probability a facility is in different star categories (for those categories where the probability is ≥ 0.10), illustrates how these data might be presented to consumers. It is clear from the bar charts that although the 45th ranked facility is a 4-star facility and the 46th a 3-star facility, based on the probabilities that the facility is in each of the star categories, the 2 facilities are hard to distinguish.

In the sensitivity analyses, we found that: (1) although point estimates of the composite measure are insensitive to the prior distribution of the "underlying" probability of developing adverse events, the probability metrics are much more sensitive. In some cases, there can be large differences when a bimodal prior instead of a uniform prior is used

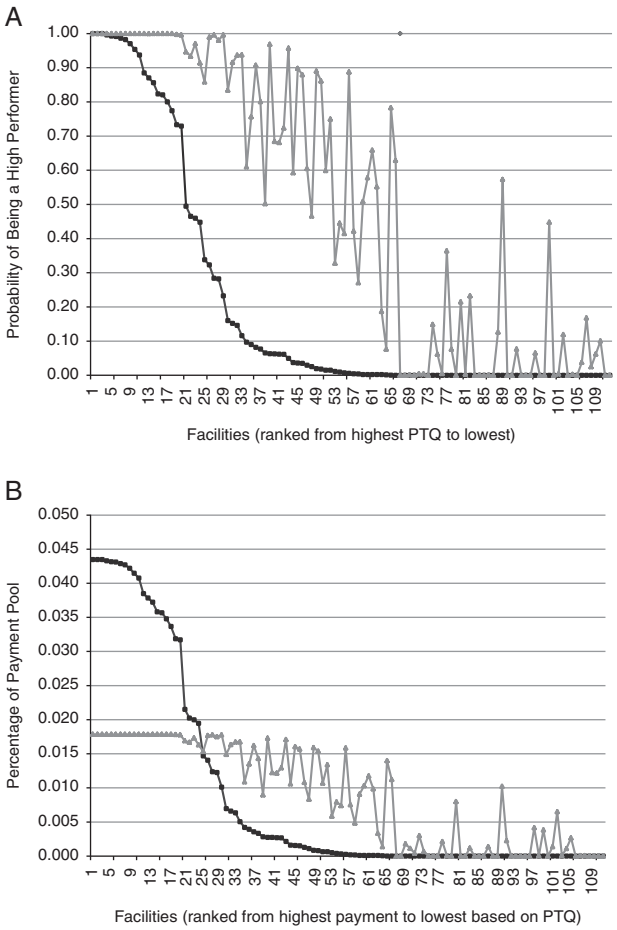

FIGURE 2. Performance and payment based on the probability a facility is in the top quintile (PTQ) and in the top half (PTH). A, Performance based on PTQ and PTH. B, Payment based on PTQ and PTH. A, The monotonically declining black line shows the probability of being a high performer when high performance is defined as being in the top quintile (PTQ); the jagged gray line shows the probability of being a high performer when high performance is defined as being in the top half of all facilities (PTH). PTQ discriminates much more among the top-ranked facilities (PTQ ranges from 1.00 to 0.46 among the top 23 ranked facilities, whereas PTH ranges from 1.00 to 0.93); PTH discriminates more among the mid-ranked facilities. B, The proportion of a pay-for-performance pool that would go to facilities if bonus payments were based on PTQ (monotonically declining black line) and PTH (jagged gray line). The greater discrimination among high performers when using PTQ is reflected in the substantially higher bonuses payments awarded to the highest of the high performers. When using PTH, mid-ranked performers receive higher bonus payments.

(Supplemental Digital Content 4, http://links.lww.com/MLR/A810, Figure 4.1); (2) The basic relation between PTQ and PTH is not sensitive to the skewness of the composite scores, although as the distribution of composite scores changes from left skewed to symmetric to right skewed, the PTQ metric results in somewhat greater discrimination among the highest of the high performers (Supplemental Digital Content 4, http://links.lww.com/MLR/A810, Figure 4.2); and

TABLE 2. Probability "Boundary" Facilities Based on the Rank of the Point Estimate of Their Composite Score are in Different Star Categories*

Probability in Indicated Star Category						
5 Stars	4 Stars	3 Stars	2 Stars	1 Star	Quintile	Rank
0.083	0.448	0.373	0.097	0.000	2	41
0.003	0.512	0.475	0.010	0.000	2	42
0.009	0.534	0.394	0.063	0.000	2	43
0.007	0.470	0.511	0.012	0.000	2	44
0.089	0.409	0.371	0.129	0.001	2	45
0.065	0.447	0.356	0.127	0.006	3	46
0.057	0.409	0.396	0.139	0.000	3	47
0.177	0.304	0.293	0.158	0.068	3	48
0.086	0.433	0.288	0.180	0.014	3	49
0.015	0.402	0.524	0.059	0.000	3	50
0.000	0.110	0.620	0.264	0.006	3	63
0.004	0.160	0.481	0.343	0.012	3	64
0.005	0.177	0.447	0.345	0.028	3	65
0.000	0.028	0.603	0.368	0.001	3	66
0.004	0.097	0.494	0.392	0.013	3	67
0.025	0.195	0.272	0.392	0.116	4	68
0.000	0.012	0.517	0.466	0.004	4	69
0.006	0.118	0.353	0.441	0.083	4	70
0.000	0.068	0.400	0.497	0.035	4	71
0.000	0.014	0.340	0.633	0.013	4	72

*The 5 facilities ranked at the bottom of quintile 2 (4 stars), the 5 ranked at the top and 5 at the bottom of quintile 3 (3 stars), and the 5 ranked at the top of quintile 4 (2 stars). For many of these facilities near a quintile "boundary," there is a substantial probability they are in a different quintile than the one assigned based on the rank of the point estimate of their composite score.

(3) Finally, in a P4P program, there are large differences in the allocation of payment bonuses to higher-performing programs when the allocation is based on applying PTQ to a composite measure versus applying PTQ to individual QIs and then aggregating the results across QIs (Supplemental Digital Content 4, http://links.lww.com/MLR/A810, Figure 4.3). When allocations are based on a composite score, the payment bonuses range from 4.3% of the pool for the highest performers to almost nothing; and when allocations are based on aggregating bonuses for individual QIs, bonuses range from 1.7% of the pool to 0.3%. When payment is based on PTH applied to individual QIs, the differences in allocation are even further compressed (1.2% of the pool to 0.6%). Not only are payments compressed, but facilities that are highest ranked when applying PTQ to the composite are often ranked substantially lower when applying PTQ to individual QIs. For example, of the 23 facilities receiving the largest bonuses based on the composite score, only 3 are in the top 23 based on individual QIs; 13 are ranked below the median. As shown in Supplemental Digital Content (http://links.lww.com/MLR/A810, Figure 4.3), there is no relationship between rankings based on the composite and rankings based on the individual QIs. Most of our QIs were not highly correlated and this may explain the lack of a relationship between the 2 approaches.

DISCUSSION

The value of a probability metric for profiling and P4P comes from being a widely understood concept (we are all now used to reporting of the probability of different weather

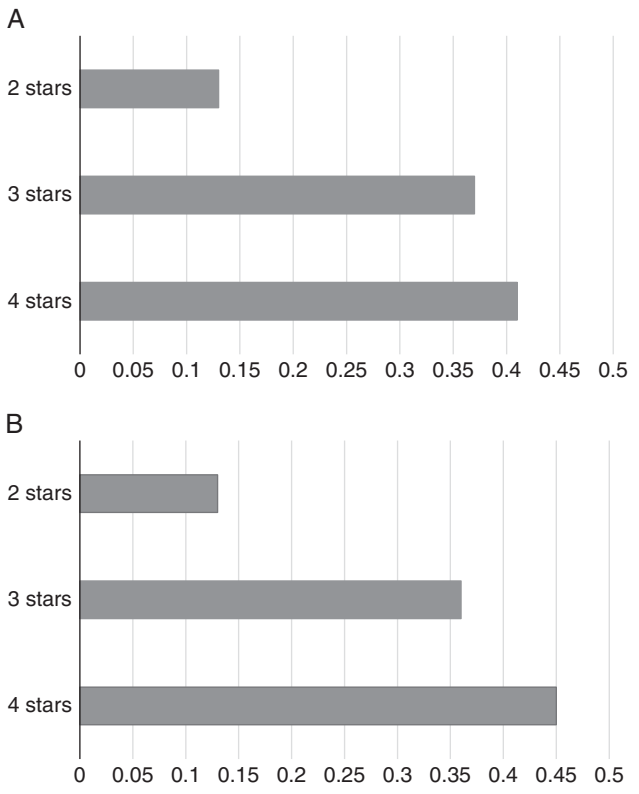

FIGURE 3. Probability a facility is in the indicated star category. A, Facility A: 4 stars (ranked 45th). B, Facility B: 3 stars (ranked 47th). These bar charts, which illustrate the way in which information might be presented to consumers, show the problem with a star classification. On the basis of the point estimate of the composite score, facility A is a 4-star facility and facility B is a 3-star facility. However, using the probability metric, facility A and B differ little in terms of being a 3-star facility (0.37 and 0.36, respectively) or a 4-star facility (0.41 and 0.45). [To simplify the presentation for consumers, we only show those star categories where there is at least a 0.10 chance the facility is in the indicated category. The higher rank of facility A is reflected in the higher probability of being a 5-star facility (0.089) compared to facility B (0.065)].

patterns) in which one number allows both ranking and communication of the likelihood that a facility is in a particular category. We have shown that our probability metric provides better discrimination of high performers than the point or interval estimate of the composite score. We also illustrated that choice of the quantile has implications for which types of facilities are distinguished in terms of their performance and the implications of this differentiation for how funds might be distributed in a P4P program. A smaller top quantile (eg, top decile or quintile) will highlight differences among top-performing facilities but provide little distinction among lower-performing facilities; a larger top quantile (eg, the top 40% or 50%) will provide little distinction among high-performing facilities but much more between lower-performing facilities. Also, in a P4P program, a smaller top quintile will allocate a higher percentage of the bonus pool to the highest-performing facilities, whereas a larger top quintile will distinguish much less between the percentage of the pool going to the highest-performing facilities compared to other facilities that are not among the very highest but still doing well. Thus, the choice of quantile depends upon the nature of the incentives one wants to create.

Although we have not considered change scores in our P4P analyses, change scores combined with use of the PTQ metric may create strong incentives for improvements among the lower-performing facilities. The highest performers in terms of performance level receive the largest payments using the PTQ metric; however, the lower payments to other facilities, aside from not providing resources for improvement efforts, may discourage mid-performing and lower-performing facilities because of the difficulty of moving into the top ranks. However, mid-performing and lower-performing facilities are the most likely to be able to significantly improve their scores. A PTQ metric used with change scores would significantly reward those who improve the most. CMS has recognized the value of rewarding both performance and improvement in its Value-Based Purchasing Program. Whatever the performance measure, stronger incentives for improvement are created by a continuous score that rewards all facilities at least at some level than by a binary score that allocates all of the payments to the top-performing facilities.

We focused on the probability that a facility is in the top quantile but one could also consider the probability a facility is in a low quantile (eg, the bottom quintile). Bonuses could be attached to the former probability and penalties to the latter. Also, a large number of organizational research studies identify high-performing and low-performing facilities and then attempt to identify structural and process characteristics that differ between the 2 types of facilities.[1–3] Better identification of high-performing and low-performing facilities through use of the probability metric may improve site selection for these types of studies. Attaching probabilities to star designations no doubt complicates the information provided to consumers. However, as illustrated in Figure 3, it provides important information for both the facility and consumer.

We illustrated the probability metric using a composite measure, whereas an approach more consistent with CMS methods would calculate the probability metric for individual QIs and then combine these into an aggregate score. As we have shown, the probability metric loses much of its ability to discriminate among facilities when it is applied to individual QIs and the results aggregated across QIs. When there are a large number of performance measures, many facilities appear to "pick and choose," concentrating their improvement efforts on certain indicators at the expense of others. It appears difficult to be among the best on some of the individual QIs and still do reasonably well on most of the other QIs. Other facilities attempt to do reasonably well on many of the QIs but at the expense of not being among the best on most of them. This same phenomenon has been shown in the context of hospitals, where facilities that do best on a composite measure are often not in the group of highest performers on many of the individual measures.[22] Although this type of analysis needs further study, it does suggest that appropriate incentives would reward both high performance on individual QIs and high performance on a composite measure.

As noted, probability metrics have not been widely used. In addition to the complexity of MCMC methods,

a more important reason is that with a large number of facilities and patients, MCMC methods can take a long time to estimate quantities of interest. Peköz et al[23] have developed an approximation that can significantly reduce the time required to estimate model parameters when using MCMC methods, which increases the feasibility of the approach.

In conclusion, we believe that continuous-valued probability-based metrics are more intuitive and understandable than current approaches used for profiling. As we have shown, this metric better distinguishes and rewards high-performing facilities than alternative approaches currently in use. However, as with all proposed measures, the continuous-valued probability-based metric needs to be presented along with competitors to various stakeholders, including consumers, to better understand its potential and applicability to real-world situations. Until a formal "vetting" has occurred, facilities may be judged unfairly based on common approaches.

REFERENCES

1. Bradley EH, Herrin J, Mattera JA, et al. Quality improvement efforts and hospital performance: rates of beta-blocker prescription after acute myocardial infarction. *Med Care*. 2005;43:282–292.
2. Vina ER, Rhew DC, Weingarten SR, et al. Relationship between organizational factors and performance among pay-for-performance hospitals. *J Gen Intern Med*. 2009;24:833–840.
3. Curry LA, Spatz E, Cherlin E, et al. What distinguished top-performing hospitals in acute myocardial infarction mortality rates? *Ann Intern Med*. 2011;154:384–390.
4. Krumholz HM, Normand S-LT, Galusha DH, et al. Risk-Adjusted Models for AMI and HF 30-Day Mortality: Methodology. 2007. Available at: http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1163010421830. Accessed January 15, 2014.
5. Centers for Medicare & Medicaid Services. National Provider Call: Hospital Value-Based Purchasing—Fiscal Year 2015 Overview for Beneficiaries, Providers, and Stakeholders. Available at: http://www.cms.gov/Outreach-and-Education/Outreach/NPC/Downloads/HospVBP_FY15_NPC_Final_03052013_508.pdf. Accessed July 9, 2013.
6. Centers for Medicare & Medicaid Services. Available at: http://www.medicare.gov/nursinghomecompare/search.html. Accessed December 10, 2013.
7. Normand S-LT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc*. 1997;92:803–814.
8. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med*. 1997;127:764–768.
9. Burgess JF Jr, Christiansen CL, Michalak SE, et al. Medical profiling: improving standards and risk adjustments using hierarchical models. *J Health Econ*. 2000;19:291–309.
10. Castle NG, Ferguson JC. What is nursing home quality and how it is measured. *Gerontologist*. 2010;50:426–442.
11. 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.
12. Kahn CN III, Ault T, Isenstein H, et al. Snapshot of hospital quality reporting and pay-for-performance under Medicare. *Health Aff*. 2006;25:148–162.
13. Edwards JR, Bagozzi RP. On the nature and direction of relationships between constructs and measures. *Psychol Med*. 2000;5:155–174.
14. Landrum MB, Bronskill SE, Normand S-LT. Analytic methods for constructing cross-sectional profiles of health care providers. *Health Serv Outc Res Method*. 2000;1:23–47.
15. Efron B, Morris CN. Stein's paradox in statistics. *Sci Am*. 1972;236:119–127.
16. Greenland S. Principles of multilevel modeling. *Int J Epidemiol*. 2000;29:158–167.
17. 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.
18. O'Brien SM, Shahian DM, DeLong ER, et al. Quality measurement in adult cardiac surgery: part 2—statistical considerations in composite measure scoring and provider rating. *Ann Thorac Surg*. 2007;83:S13–S26.
19. Staiger DO, Dimick JB, Baser O, et al. Empirically derived composite measures of surgical performance. *Med Care*. 2009;47:226–233.
20. Shwartz M, Peköz EA, Christiansen CL, et al. Shrinkage estimators for a composite measure of quality conceptualized as a formative construct. *Health Serv Res*. 2013;48:271–289.
21. Rosen AK, Chen Q, Borzecki AM, et al. Using estimated true safety event rates vs. flagged safety event rates: does it change hospital profiling and payment? *Health Serv Res*. 2014;49:1426–1445.
22. Shwartz M, Cohen AB, Restuccia JD, et al. How well can we identify the high-performing hospital? *Med Care Res Rev*. 2009;68:290–310.
23. Peköz EA, Shwartz M, Christiansen CL, et al. Approximate models for aggregate data when individual-level data sets are very large or unavailable. *Stat Med*. 2010;29:2180–2193.