Hospital Size, Uncertainty, and Pay-for-Performance

Gestur Davidson, Ph.D., Ira Moscovice, Ph.D., and Denise Remus, Ph.D., R.N.

We construct statistical models to assess whether hospital size will impact the ability to identify "true" hospital ranks in pay-for-performance (P4P) programs. We use Bayesian hierarchical models to estimate the uncertainty associated with the ranking of hospitals by their raw composite score values for three medical conditions: acute myocardial infarction (AMI), heart failure (HF), and community acquired pneumonia (PN). The results indicate a dramatic inverse relationship between the size of the hospital and its expected range of ranking positions for its true or stabilized mean rank. The smallest hospitals among the augmented dataset would likely experience five to seven times more uncertainty concerning their true ranks.

INTRODUCTION

The HQID, sponsored by CMS, is being conducted with participating hospitals that are members of Premier Inc., a national alliance of non-profit hospitals. It was designed to examine whether a system that explicitly pays-for-performance (establishing rewards for high performance and penalties for low performance) can lead to systemwide improvement in the quality-of-care that hospitals provide in selected medical conditions.

In the HQID, high performance and some aspects of low performance are defined in a relative sense through the ranking of hospitals' performance in providing specific services. Because hospitals vary greatly in the annual number of patients seen with the medical conditions included in the HQID, statistical theory suggests that smaller hospitals can expect to experience much greater sampling variability in their performance scores.

In this study we construct Bayesian statistical models to assess the impact that hospital size is likely to have on the ability to infer true ranks in P4P programs patterned after HQID. Specifically, we address the following questions:

- How accurately can we expect to predict true performance scores for hospitals participating in HQID?
- What unintended consequences might arise in a system that establishes rewards and penalties based on relative performance when measured with substantial variation in accuracy?

BACKGROUND

The demonstration is a 3-year project with data collected from participating hospitals from October 1, 2003 - September 30, 2004 (Year 1), October 1, 2004 - September 30, 2005 (Year 2), and October 1, 2005 - September 30, 2006 (Year 3).

The demonstration measures quality in five clinical areas: (1) AMI, (2) coronary artery bypass graft procedures, (3) HF, (4) PN, and (5) hip and knee replacement procedures. However, because coronary artery bypass graft and hip and knee

NOTE: The authors used (with approval from CMS) the Hospital Quality Incentive Demonstration (HQID) Year-1 data for their research.

Gestur Davidson and Ira Moscovice are with the University of Minnesota and Denise Remus is with BayCare Health System. The research in this article was supported by the Public Health Service under Grant Number 5U1CRH03717-02-00. The statements expressed in this article are those of the authors and do not necessarily reflect the views or policies of the University of Minnesota, BayCare Health System, or the Centers for Medicare & Medicaid Services (CMS).

Table 1
Associated Reporting Measures Used in the Hospital Quality Incentive Demonstration, by Clinical Conditions

Clinical Condition	Measures
Acute Myocardial Infarction	Aspirin at Arrival[1,2,3,4] Aspirin Prescribed at Discharge[1,2,3,4] ACEI for LVSD[1,2,3,4] Smoking Cessation Advice/Counseling[1,2,3] Beta Blocker Prescribed at Discharge[1,2,3,4] Beta Blocker at Arrival[1,2,3,4] Thrombolytic Received Within 30 Minutes of Hospital Arrival[1,5,7] PCI Received Within 120 Minutes of Hospital Arrival[1,5,7] Inpatient Mortality Rate[1,3,6,8]
Heart Failure	Left Ventricular Function Assessment[1,2,3,4] Detailed Discharge Instructions[1,2,3] ACEI for LVSD[1,2,3,4] Smoking Cessation Advice/Counseling[1,2,3]
Community Acquired Pneumonia	Percentage of Patients Who Received an Oxygenation Assessment Within 24 Hours Prior to or after Hospital Arrival[1,2,3,4] Pneumococcal Screening/Vaccination[1,2,3,4] Blood Culture Collected Prior to First Antibiotic Administration[1,2,3] Smoking Cessation Advice/Counseling[1,2,3] Initial Antibiotic Consistent with Current Recommendations[1,2,7] Influenza Screening/Vaccination[1,2,7] Antibiotic Timing – Percentage of Pneumonia Patients Who Receive First Dose of Antibiotics Within 4 Hours After Hospital Arrival[1,2,4,7]

[1] National Quality Forum.

[2] Centers for Medicare & Medicaid Services (CMS) 7th Scope of Work.

[3] Joint Commission on Accreditation of Health Care Organizations (JCAHO) Core.

[4] The National Voluntary Hospital Reporting Initiative (AHA Initiative).

[5] The Leapfrog Group (proposed).

[6] Risk adjusted using JCAHO methodology.

[7] CMS and/or JCAHO to align with this measure in 2004.

[8] Outcome measure (all other measures in Table 1 are process measures).

NOTES: ACEI is angiotension-converting enzyme inhibitor. LVSD is left ventricular/systolic dysfunction. PCI is percutaneous coronary intervention.

SOURCE: Centers for Medicare & Medicaid Services: Premier Hospital Quality Incentive Demonstrations, Terms and Conditions, 2006.

replacement procedures are performed so infrequently in smaller, rural hospitals, this study includes only AMI, HF, and PN.

The HQID includes financial incentives for high quality in each of the 3 years and, beginning in the third year, applies financial penalties for scores that fall below low quality thresholds based on the distribution of scores in the first year of the demonstration.

Table 1 presents each of the three medical conditions examined in the present analysis and their associated reporting measures, as used in the HQID. Using PN as an example, a hospital receives a composite condition score (based on patients with PN that the hospital treated during the year) that is computed as the ratio of the total number of services that were received by those patients to the total number of services deemed to be needed by those patients. We use the term number of services needed by patients to refer to the number of patients with a condition who were determined by providers to require the service associated with the quality measure. The quality measures implemented, and the calculation of composite scores, exclude patients transferred from or to another acute care hospital.

Rewards and Penalties

There are two types of HQID rewards: (1) financial incentives (increased payments), and (2) publicly announced placement in the top 50 percent of the hospitals for each clinical condition used in the demonstration.

Composite scores are calculated for each clinical condition from the reported data of all participating Premier Inc. hospitals in Year 1, with the number of hospitals varying by condition. The composite scores for the hospitals are then ranked, and each hospital's percentile is determined.

A hospital with a composite clinical quality score in Year 1 that places it in the top decile (10th percentile or higher) receives a bonus of 2 percent of the diagnosis related group-based prospective payment for the patients with the condition among all Medicare fee-for-service beneficiaries in Year 1 (Centers for Medicare & Medicaid Services, 2006). Hospitals with composite scores that place them in the second decile receive 1 percent added to their Medicare payment for that condition.

The penalties (referred to as payment adjustments) are determined as follows:

- A hospital with a composite clinical condition score in Year 3 that is below the 90th percentile cutoff composite score for that clinical condition in Year 1 will have 2 percent deducted from their Medicare payment for that condition in Year 3.
- A hospital with a composite clinical condition score in Year 3 that is below the 80th percentile cutoff composite score, but above the 90th percentile cutoff composite score, from Year 1 will have 1 percent deducted from their Medicare payment for that condition in Year 3 (Centers for Medicare & Medicaid Services, 2005a).

Finally, on the CMS Web site there is a public announcement that identifies alphabetically hospitals that have placed in the top 50 percent of that year's ranking for each clinical condition. Hospitals that had composite scores below the 50th percentile (the median) will not be named/identified on the Web site (Centers for Medicare & Medicaid Services, 2005b). This announcement can be seen as both rewarding and as a penalty and could be an important public relations consideration for hospitals not in the top 50 percent of the rankings.

VARIABILITY IN COMPOSITE SCORES

All participating Premier Inc. hospitals had composite scores that fell below 100 percent in Year 1 for each of the three conditions. Some group of factors must account for the lack of perfect provision of these indicated services. Moreover, there is a substantial range in composite scores across hospitals. Clearly there must be variability across hospitals in the relevant number and/or the relative impact of the factors causing these shortfalls.

Composite scores can be calculated as a weighted average of the individual success rates of providing each measure's service weighted by each measure's share of the total needed services for the condition. A hospital could maintain the same success rates for all the indicated services over time and yet still have substantial variability in its overall composite score for that condition if there were variability in the shares of total needed services. Further, service specific success rates within a given medical condition might themselves vary due to simple sampling variability and one would expect variation in scores arising solely due to different number of patients in a year.

Throughout this article we reference true hospital rank. By this, we mean the rank the hospital would have achieved in the steady-state if we could repeat the experiment of conducting the first year of the HQID data many times. This study involves conceiving of a hospital's composite quality score as an estimate of its steady-state score. Consequently, this study conceives of a hospital's rank/percentile as also being an estimate of its steady-state rank/percentile. We empirically implement this concept of true or steady-state composite quality scores and ranks through the use of Bayesian hierarchical models.

HOSPITAL RANKINGS

There is a growing number of statistical analyses that demonstrate the difficulty of achieving policy-relevant estimates of ranks/percentiles due to the varying size of the samples that they are based on (Lockwood, Louis, and McCaffrey, 2002; Lin et al., 2004; Marshall and Spiegelhalter, 1998; Andersson, Carling, and Mattson, 1998; Goldstein and Spiegelhalter, 1996; Normand, Glickman, and Gatsonis, 1997).

Marshall and Spiegelhalter (1998), based on their statistical methods for assessing the reliability of these ranks, conclude:

- Institutional ranks are extremely unreliable statistical summaries of performance.
- Institutions with smaller numbers of cases may be unjustifiably penalized or credited in comparison exercises.
- Additional statistical analysis may help to identify the few institutions worthy of review.
- Any performance indicator should always have an associated statistical sampling variability.

For the analysis we use a Bayesian, hierarchical modeling strategy to estimate the uncertainty associated with the ranking of hospitals by their raw composite score values. As noted by Lockwood et al. (2002), the Bayesian perspective "… provides an integrated, coherent structure in which to evaluate ranking procedures."

Bayesian Modeling

For the two conditions of HF and PN, and for the eight process measures available for AMI, our Bayesian models all share a common structure. Using HF as an example, for each of the four process measures (Table 1), we assume as fixed and known the net number of cases (NNC[i]) for each of the sampled hospitals (i.e., the number of patients seen that year in the hospital with HF who were not transfers into or out of the hospital). We then posit that for each hospital [i from 1 to 265], the NNC[i] needing each of the four measures services is distributed as a binomial, with binomial parameter needp[i]. Similarly, we posit that among the patients needing each measure's service, the number who receive it are also binomially distributed, with parameter recvp[i].

Bayesian models were used to obtain each hospital's true or steady-state values of these two sets of binomial parameters: (1) the proportion of the net cases needing each measure's service and (2) the proportion of those needing each service receiving it.

Using a Bayesian framework, we assume that the 265 hospitals share a distribution for each one of the eight binomial parameters. We assume that the true values of each needp[i] binomial parameter for the 265 hospitals all come from a common distribution, meaning that together their 265 true values come from a distribution that has a mean, and their distinctness is reflected in the variance of this common distribution. We do not know the true values of the mean or the variance of this common distribution, but we have a general idea of their range. We posit some prior knowledge for them in the form of an assumption about the distribution from which these parameters in turn are likely to be drawn.

Through computer intensive sampling iterations, Bayesian models allow us to derive estimates of the values of the parameters from each of the modeling levels. Once convinced the models have converged to steady-state values, we carefully inspect these values to see if they have converged to plausible values. Finally, we test, to be sure that the values obtained in the model for the parameters of interest are not dependent on the assumptions employed in the models.

When convinced the models have converged, we allow the sampling algorithm to continue to run to trace out the full distribution (the posterior distribution) of each hospital's values of the needp[i] and recvp[i] binomial parameters. Specifically, the full Bayesian model provides the following output:

- The values of the eight binomial parameters for each hospital for each iteration of the post-convergence simulation are inserted into the composite score formula. We generate the posterior distribution of the composite scores for each hospital.
- From this posterior distribution of the composite scores of each hospital we take its mean as the estimate of each hospital's true or steady-state value of its HF composite score.
- In each of the iteration of the post-convergence simulation, the individual hospital composite scores are used to compute the rank of each hospital. Thus, we generate for each hospital a posterior distribution of its ranks.
- From this posterior distribution of ranks we obtain a mean rank of each hospital for HF, which is an estimate of the true or steady-state value of its rank.
- The full posterior distribution of each hospital's ranks are obtained and from this we directly assess the range of each hospital's ranks. In particular, we can readily show the 95 percent confidence intervals (CIs) for the mean ranks—or 95 percent credible intervals in Bayesian modeling. These 95 percent credible intervals for the mean ranks constitute our primary metric for the amount of uncertainty inherent in estimated ranks, which the cited literature strongly recommends be a part of any presentation of ranks.

This rank estimator is optimal for ranks in the general sense of providing us with the best estimates for all ranking positions/percentiles. If there is interest in whether a hospital has a rank placing it in the top 20 percentile or below it, then this overall rank estimator is not optimal (Lin et al., 2003). Because rewards in the HQID are based on just such specific ranking thresholds, a second ranking estimator is derived. Specifically, from the posterior distribution of ranks we count the number of times (i.e., post-convergence iterations) each hospital's estimated rank exceeds or falls below the percentile cut point of interest. Over all these post-convergence iterations, this yields an estimated probability of exceeding or falling below this percentile cut point.

For the AMI condition, in addition to the eight process measures there is a ninth outcome measure, the standardized survival ratio (SSR). A separate Bayesian hierarchical model is used to estimate the stabilized values of this ratio for the sample of hospitals that is adapted from the Bayesian model used by Liu et al. (2003) in their study of mortality within dialysis centers. We combine this SSR component with the composite score calculated from the eight AMI process measures using the formula provided by CMS. Specifically, the overall composite score for the AMI patients for a hospital is a weighted average—89 percent of the composite score from the eight process measures plus 11 percent of its SSR value.

DATA SOURCES

We used the HQID data for Year 1. The identity of the hospitals associated with the reporting items was masked to us, and no individual patient-level data were used. For the quality measures in each condition (AMI, HF, PN), the reporting items were:

- Total number of patients seen in the hospital in Year 1.
- Number of the total that were transfers, either into or out of the hospital.
- Number of the non-transfer patients needing the service associated with the measure.
- Number of the non-transfer patients needing the service associated with the measure who received the service.
- For the specific case of AMI, the expected mortality rate associated with the AMI non-transfer patients seen by that hospital and the number of actual deaths.

We received each hospital's number of licensed and staffed beds; however, because the data on beds were incomplete, we present all empirical results by our size metric: hospitals' (NNC) for each medical condition.

Hospital Compare Program Data

The hospitals participating in the HQID are not representative of the full population of short-term, general hospitals in the U.S. They include only 3 community access hospitals (CAHs), and 44 rural hospitals. But CAHs constitute 23 percent of all short-term, general hospitals. Since the goal of this project is to specifically show the influence of small hospital size per se on the likely variability in hospital ranks, we undertook a second set of model runs using additional data from the Hospital Compare for voluntarily participating hospitals (Casey and Moscovice, 2006). The measures for HF and PN in Hospital Compare are identical to those we are using from the HQID.

For each of the three medical conditions, we have drawn a random sample of the data provided by Hospital Compare participants that are CAHs.[1] The proportion of CAHs in the total augmented dataset (Premier Inc. hospitals plus Hospital Compare CAHs) was set equal to the proportion of CAHs (23 percent) in the population of short-term, general hospitals.

There are some complications in using the Hospital Compare data. First, the Hospital Compare data include all of the eight AMI process measures, but neither AMI deaths nor the expected AMI mortality rate. Thus, the SSR could not be computed using these data. However, using only the Premier Inc. dataset on 243 hospitals, the ranks of hospitals determined from the full AMI composite scores and the ranks determined from just the AMI process composite scores were found to be virtually identical, as was the variability in the mean ranks. Adding the SSR to the AMI process composite scores within the models does not yield meaningful differences for the purpose of this study. Thus, we use and report the results of the AMI model from the augmented dataset.

Secondly, there were differences in the duration of the collection periods for some of the measures in the Hospital Compare data. The Hospital Compare data had a starter set of measures that were later augmented by additional measures.

These augmented measures were reported for less than or equal to 9 months. Consequently, we normalized their values to represent full year measures.[2] We feel confident that this adjustment for the augmented measures had no meaningful impacts on our results. Specifically, we carried out analyses both with the Premier hospital only—where all measures were reported for a full year—and then with Premier hospitals augmented by year. For

[1] Depending on the condition and measure, as many as 468 CAHs reported data in the Hospital Compare for the period of interest.

[2] For the augmented measures we used an adjustment factor of 4/3 against all reported data.

Table 2

Percent Distribution of Composite Scores for Community Required Pneumonia (PN), Heart Failure (HF), and Acute Myocardial Infarction (AMI)

	Composite Scores		
Percentile	PN	HF	AMI
		Percent	
10th	66.4	51.1	79.4
25th	71.0	59.8	85.6
Median	76.4	69.6	89.9
75th	82.1	80.2	93.5
90th	86.0	86.1	95.7
Lowest	57.0	25.4	49.0
Highest	92.4	96.4	99.4
Percentage of Hospitals With Composite Scores ±2 Composite Points Around the Median	23.0	9.0	29.0

NOTE: Distribution of composite scores is for Premier Inc. hospitals only.

SOURCE: Davidson, G. and Moscovice, I., University of Minnesota, and Remus, D., BayCare Health System: Analysis of year-1 data from the Health Quality Incentive Demonstration, 2007.

the second smallest size strata that drew meaningfully from both sets of data, we observed no meaningful differences in the average widths of the 95 percent CIs from these two sets of Premier-only and Hospital Compare-augmented data.

EMPIRICAL FINDINGS

We begin with descriptive statistics on the distribution of composite scores for each of the three conditions. The degree to which hospitals are closely clustered together in their composite scores will impact the performance of any ranking procedure employed, with greater or lesser effects depending on the amount of measurement error due to sampling variability for the composite scores. Table 2 provides composite scores associated with hospitals at the 10th, 25th, median, 75th, and 90th percentiles for each condition and the percent of hospitals that are found within a band of ± 2 composite score points around the median value for each condition.

From the PN composite scores for the 263 hospitals we observe a significant degree of clustering. Fully 50 percent of the sample hospitals (from 25th to 75th percentiles) have composite scores falling within a range of 11 composite score points (between 71 and 82) and 80 percent of the hospitals (from 10th to 90th percentiles) have composite scores within a range slightly less than 20 composite score points (between 66.4 and 86). Finally, 23 percent of the hospitals are found within a band of ± 2 composite points around the median value of 76.4 for PN.

For the HF composite scores it takes a range of 26 composite points to include the middle 50 percent of the sample hospitals, and to enclose the middle 80 percent of the hospitals a range of 36 composite scores points is necessary. Only 9 percent of the hospitals are within a band of ± 2 composite points around the median value of 69.6 for HF.

Lastly, AMI composite scores, based on 262 hospitals[3] exhibit the greatest amount of clustering. For AMI it takes only a range of 8 composite points to include the middle 50 percent of the sample hospitals, considerably less than the 11 composite points for PN. To enclose the middle 80 percent of the hospitals, a range of only 16.4 composite scores points is needed. Twenty-nine percent of the hospitals are within a band of ± 2 composite points around the median value (89.9).

[3] We use the full 262 count of hospitals for this exercise, not the smaller number feasible for modeling.

Table 3

Number of Hospital Community Acquired Pneumonia (PN) Patients Per Year on the Width of the 95 Percent Confidence Intervals (CIs) for Hospital Ranks and Percentile Values

	Average Range of	
PN Patients in Hospital	Rank Positions Falling Within 95 Percent CI for Ranks	Percentile-Points Falling Within 95 Percent CI for Ranks
≤ 20	221	64
21–40	168	49
41–60	134	39
61–100	121	35
101–150	80	23
151–200	71	21
201–250	77	22
251–300	64	19
301–350	65	19
351–400	62	18
401–450	69	20
451–500	60	17
501–600	56	16
601–700	44	13
701–800	44	13
801–900	48	14
901–1000	46	13
1,001–1,100	38	11
1,101– 2,313	35	10

NOTE: Premier Inc. hospitals plus community access hospitals sample: *n*=344.

SOURCE: Davidson, G. and Moscovice, I., University of Minnesota, and Remus, D., BayCare Health System: Analysis of year-1 data from the Health Quality Incentive Demonstration, 2007.

Uncertainty About True Hospital Ranks

There are a number of ways to portray the amount of uncertainty in estimates of the true relative performance of hospitals. Our main metric for portraying this uncertainty is the 95 percent CIs about the mean rank derived from the Bayesian models for each medical condition. To show how small size increases the expected amount of uncertainty, we stratify the entire sample into 19 size strata and give the average width of the 95 percent CI of the hospitals in each stratum (the average number of ranking positions between the upper 95 percent CI rank value and the lower 95 percent CI rank value). We also provide the translation of ranking positions into the equivalent range of percentile points, which directly expresses the degree of uncertainty in true performance relative to the entire 100 percentile-point range. To facilitate comparisons across the three conditions, we use the same size groupings across the three medical conditions although the distributions of net cases varies somewhat.

We provide and discuss only the model results derived from using the augmented sample of HQID participating hospitals plus the CAHs obtained from Hospital Compare. There are two justifications for doing so: (1) this augmented sample with its additional set of smaller hospitals more completely illustrates the relationship between uncertainty and hospital size; and (2) for the hospitals larger than the smallest ones that this augmented sample introduces, the implications for uncertainty are the same in the two samples.

Table 3 illustrates, for PN patients, the dramatic inverse relationship between the size of the hospital and the expected range of ranking positions about its true or stabilized mean rank. For the smallest hospitals (20 or less PN patients per year) the average range of ranking positions is 221 out

Table 4

Number of Hospital Heart Failure Patients (HF) Per Year on the Width of the 95 Percent Confidence Intervals (CIs) for Hospital Ranks and Percentile Values

	Average Range of	
HF Patients in Hospital	Rank Positions Falling Within 95 Percent CI for Ranks	Percentile-Points Falling Within 95 Percent CI for Ranks
≤ 20	161	46
21–40	112	32
41–60	90	26
61–100	85	24
101–150	66	19
151–200	56	16
201–250	53	15
251–300	53	15
301–350	40	12
351–400	46	13
401–450	38	11
451–500	37	10
501–600	34	10
601–700	28	8
701–800	29	8
801–900	26	7
901–1000	27	8
1,001–1,100	28	8
1,101–1,926	25	7

NOTE: Premier Inc. hospitals plus community access hospitals sample: *n*=348.

SOURCE: Davidson, G. and Moscovice, I., University of Minnesota, and Remus, D., BayCare Health System: Analysis of year-1 data from the Health Quality Incentive Demonstration, 2007.

of the 344 hospitals in this sample, or a full 64 percentile points. At nearly two-thirds of the entire range of percentiles, this represents substantial uncertainty about the measurement of true relative performance of the smallest hospitals.

For the largest size stratum (more than 1,100 PN patients per year) this uncertainty extends to only 35 ranking positions, or 10 percentile points. We conclude from this exercise that for PN patients that the smallest hospitals would likely experience—through the use of ranks of annual composite scores—approximately six times more uncertainty about their true ranking positions than the largest hospitals. Also of interest is the relatively large number of PN patients needed to achieve even a 20-percentile range in their true score, on average.

For HF there is also a strong inverse relationship between the size of the hospital and the expected range of ranking positions for its true or stabilized mean ranks (Table 4). For hospitals with 20 or fewer HF patients per year the average width of the 95 percent CI for Bayesian ranks is 161 ranking positions out of 348 hospitals, or 46 percentile points. This is considerably less than the average range of ranking positions of 64 percentile points for PN patients for this size stratum. For the largest size stratum (more than 1,100 HF patients per year) this uncertainty drops to 25 ranking positions, or 7 percentile points. We conclude that there would be less uncertainty in hospitals' estimated ranks for HF than PN. Comparing the average width of the 95 percent CI for the smallest to the largest hospital size category, however, still yields roughly six times more uncertainty for the smallest hospitals compared to the largest ones.

For any given patient-size category, the reduction in uncertainty concerning true relative performance in HF compared to PN would be predicted from the differences in the distribution of composite

Table 5

Number of Hospital Acute Myocardial Infarction (AMI) Patients Per Year on the Width of the 95 Percent Confidence Intervals (CIs) for Hospital Ranks and Percentile Values

	Average Range of	
AMI Patients in Hospital	Positions Falling Within 95 Percent CI for Ranks	Percentile-Points Falling Within 95 Percent CI for Ranks
≤ 20	199	63
21–40	157	50
41–60	127	40
61–100	122	39
101–150	97	31
151–200	80	25
201–250	89	28
251–300	62	20
301–350	62	20
351–400	74	23
401–450	64	20
451–500	51	16
501–600	55	17
601–700	43	14
701–800	48	15
801–900	44	14
901–1,000	40	13
1,001–1,100	43	14
1,101–1,926	29	9

NOTE: Premier Inc. hospitals plus community access hospitals sample: *n*=314.

SOURCE: Davidson, G. and Moscovice, I., University of Minnesota, and Remus, D., BayCare Health System: Analysis of year-1 data from the Health Quality Incentive Demonstration, 2007.

scores for the two conditions as provided in Table 2, since the distribution of HF composite scores was spread out much more than was the case for PN.[4] A less concentrated distribution of composite scores for HF is the equivalent of a stronger signal, or more information about true relative performance.

For AMI (Table 5), we would expect to see the greatest amount of uncertainty displayed in true ranks based on the results of Table 2, and for the most part we do. For 20 or fewer AMI patients per year, the average width of the 95 percent CI for Bayesian ranks is 199 ranking positions out of 314 hospitals, or a range that represents 63 percentile points, comparable to that observed for PN patients. Using the ratio of the smallest to largest size stratum, there is roughly seven times more uncertainty for the smallest hospitals compared to the largest ones concerning their true rank.

Hospital Placement

We summarize the uncertainty of hospital placement in the top 20 percent in Table 6 since this is the specific way that ranks are used in the HQID for assigning rewards. The following are measures of uncertainty:

- Hospitals are ranked by their Bayesian model probabilities of being in the top 2 deciles (20 percentile or better) with 95 percent or greater probability.
- The share of hospitals assigned to be in the top 20 percent of hospitals that have Bayesian model probabilities of being in the top 20 percent with less than 80 percent probability. We chose the 80 percent benchmark-level since if not the ideal, it is at least a reasonable level of probability for assigning the last hospital in the top 20 percent.
- For the group of hospitals previously identified, the average probability of being in the top 20 percent. This measure reflects how quickly or slowly

[4] The distribution of composite scores in Table 2 is for Premier Inc. hospitals only.

Table 6

Measures of Uncertainty Concerning Top 20 Percent Placement in Rank: Premier Inc. Hospital Plus Community Access Hospital Sample

Ranking	Community Acquired Pneumonia	Heart Failure	Acute Myocardial Infarction
		Percent	
Percentage of Hospitals Placed in Top 20 Percent That Have 95 Percent or Greater Probability of Being in Top 20 Percent	52	57	49
Percentage of Hospitals Placed in Top 20 Percent That Have Less Than 80 Percent Probability of Being in Top 20 Percent	33	23	37
Average Probability of Being in Top 20 Percent for the Hospitals That Have Less Than 80 Percent Probability of Being in Top 20 Percent	61	67	57
Probability of Being in Top 20 Percent for the Last Hospital Assigned to the Top 20 Percent	43	55	43

SOURCE: Davidson, G. and Moscovice, I., University of Minnesota, and Remus, D., BayCare Health System: Analysis of year-1 data from the Health Quality Incentive Demonstration, 2007.

these probabilities of being in the top 20 percent decline.

- The probability of being in the top 20 percent of hospitals for the very last hospital that makes the top 20 percent list. This reflects how uncertain we are at the margin, for the last hospital that is in the top 20 percent.

Although there is some variation in the measures across the three conditions, an important policy conclusion is the low level of confidence that we have for many of the hospitals that would be assigned to the top 20 percent of hospitals by virtue of having the highest probabilities of possessing true ranks that justify that position. Specifically, for only 49 to 57 percent of the hospitals assigned to the top 20 percent would this placement have the conventional 95 percent confidence or higher. Looking at the other end of the top 20 percent group, from 23 to 37 percent of those assigned to the top 20 percent would have probabilities of less than 80 percent that their true ranks justified that placement, and the average of these probabilities is quite low, between 57 to 67 percent. This reflects the sharp drop off in the probabilities of being in the top 20 percent below the 80 percent benchmark level. This lower end is also reflected in the low probabilities of the last hospitals assigned to the top 20 percent, as low as 43 percent for PN and AMI.

DISCUSSION AND POLICY IMPLICATIONS

From these results, we identify the following major take away points that are important for policy arising from a P4P system like the HQID that defines quality through the use of simple ranks of composite scores.

- A clear message found in all the literature is the necessity of accompanying estimates of rank/percentile placement with adequate measures of the uncertainty of those estimates. This is good statistical practice and essential to the crafting and conduct of good policy.
- Identifying relative quality from simple ranks based on annual composite scores will impact smaller institutions to a greater extent than larger institutions. Smaller hospitals have increased likelihoods of placing in and out of the

top 20 percentile of ranks that defines and rewards highest quality and the top 50 percent that would bring public recognition.

- The likelihood and consequences of high levels of uncertainty concerning hospitals' relative levels of quality differs by specific medical condition, but in all cases it would be large enough to have important implications for policy.
- The findings are likely to be generalizable to hospitals beyond this sample. While both the Premier Inc. hospital sample and the augmented sample are not random draws of all hospitals in the country, there is no reason to believe a priori that the results from other samples of hospitals would differ in any policy meaningful way.
- The results based on the first year of the HQID are likely to understate the degree of uncertainty that would be associated with more mature P4P programs that use rankings like the HQID. The natural evolution of any reasonably successful P4P program (borne out by preliminary data from Years 2 and 3 of the HQID) would likely lead to increased concentration of scores over time. However, the measurement error associated with composite scores would not decrease with this higher concentration. With greater concentration of these composite scores, the difference between higher and lower scores would increasingly be dominated by this measurement error, leading to substantial increase in the uncertainty about whether differences in annually observed ranks reflect differences in true quality scores.
- With these Bayesian models, we have begun to address the policy relevant issue of identifying and estimating the likely amount of uncertainty inherent in measuring relative quality. What is needed for good policymaking is the identification of ways to both accurately identify that uncertainty and appropriately integrate these assessments within P4P reward and recognition systems. A number of approaches to the identification, reduction, and effective management of uncertainty can be acknowledged. These include combining all measures into a single, hospital-wide composite score; the exclusive use of the composite score metric; aggregating data over longer time periods for smaller hospitals; aggregating data over a number of small hospitals; and integrating uncertainty directly into reward algorithms and public reporting. What will be critical to the success of P4P programs is the careful conceptual and empirical assessment of the benefits and limitations of various ways of scoring quality, including the relative amount of uncertainty associated with them.

ACKNOWLEDGMENTS

The authors would like to thank Brad Carlin for his help with the development of the basic Bayesian model and Michelle Casey and Patricia Bland for their help with the Hospital Compare dataset.

REFERENCES

Andersson, J., Carling, K., and Mattson, S.: Random Ranking of Hospitals is Unsound. *Chance* 11(3):34-37, 39, Summer 1998.

Carlin, B. and Louis, T.: *Bayes and Empirical Bayes Methods for Data Analysis,* 2nd Edition. Chapman and Hall/CRC Press. Boca Raton, FL. 2000.

Casey, M. and Moscovice, I.: *CAH Participation and Initial Results of Hospital Compare Data. Flex Monitoring Team Briefing Paper Number 9.* Upper Midwest Rural Health Research Center. University of Minnesota, Minneapolis, MN. February 2006.

Congdon, P.: *Applied Bayesian Modeling.* John Wiley & Sons, Ltd. Chichester, England. 2003.

Centers for Medicare & Medicaid Services: *Premier Hospital Quality Incentive Demonstration, Fact Sheet*. 2005a. Internet address: http://www.cms.hhs.gov/HospitalPremierFS200602.pdf (Accessed 2007.)

Centers for Medicare & Medicaid Services: *Premier Hospital Quality Incentive Demonstration, Historical Data FAQs*. 2005b. Internet address: http://www.cms.hhs.gov/HospitalQualityInits/downloads/HospitalHistoricalData/FAQ.pdf (Accessed 2007.)

Centers for Medicare & Medicaid Services: *Premier Hospital Quality Incentive Demonstration, Terms and Conditions*. 2006. Internet address: http://www.cms.hhs.gov/HospitalQualityInits/downloads/HospitalTermsAndConditions200601.pdf (Accessed 2007.)

Gilks, W., Richardson, S., and Spiegelhalter, D.: *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC Press. Boca Raton, Florida. 2000.

Goldstein, H. and Spiegelhalter, D.: League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society, Series A Statistics in Society* 159(3):385-443, 1996.

Lin, R., Louis, T., Paddock, S., et al.: *Ranking USRDS Provider-Specific SMRs from 1998-2001*. Johns Hopkins Department of Biostatistics Working Paper Number 67, December 2004. Internet address: http://www.bepress.com/jhubiostat/paper67 (Accessed 2007.)

Lockwood, J., Louis, T., and McCaffrey, D.: Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability Systems. *Journal of Educational and Behavioral Statistics* 27(3):255-270, Fall 2002.

Louis, T. and Shen, W.: Innovations in Bayes and Empirical Bayes Methods: Estimating Parameters, Populations and Ranks. *Statistics in Medicine* 18(17-18):2493-2505, 1999.

Marshall, E. and Spiegelhalter, D.: The Reliability of League Tables of In Vitro Fertilization Clinics: Retrospective Analysis of Live Births. *British Medical Journal* 316(7146):1701-1705, June 6, 1998.

Normand, S., Glickman, M., and Gatsonis, C.: Statistical Methods for Profiling Providers of Medical Care: Issues and Applications. *Journal of the American Statistical Association* 92(439):803-814, September 1997.

Shen, W. and Louis, T.: Triple-Goal Estimates in Two-Stage Hierarchical Models. *Journal of the Royal Statistical Society, Series B Statistical Methodology* 60(2):455-471, June 1998.

Spiegelhalter, D., Thomas, A., Best, N., et al.: *WinBUGS User Manual*. Version 1.4, January 2003.

Reprint Requests: Gestur Davidson, Ph.D., University of Minnesota, 2221 University Avenue, SE, Suite-112, Minneapolis, MN 55414. E-mail: david064@umn.edu