Upper Midwest
Rural Health
Research Center

www.uppermidwestrhrc.org

Ira Moscovice, PhD, Director
Mary Wakefield, PhD, Deputy Director

Hospital Size, Uncertainty and Pay-for-Performance

Working Paper Series — #3

Gestur Davidson, PhD, University of Minnesota
Ira Moscovice, PhD, University of Minnesota
Denise Remus, PhD, RN, Formerly, Premier, Inc.

February 2007

RHRC
Rural Health Research & Policy Centers
Funded by the Federal Office of Rural Health Policy
www.ruralhealthresearch.org

The Upper Midwest Rural Health Research Center is a partnership of the University of Minnesota Rural Health Research Center and the University of North Dakota Center for Rural Health

Acknowledgements

The authors thank Brad Carlin, Professor, Division of Biostatistics, School of Public Health, University of Minnesota, for help with the development of the basic Bayesian model used in the analysis. He is not responsible for any errors in—nor should he be construed as endorsing—the findings of this study. This study could not have been undertaken without hospital data from the CMS/Premier Hospital Quality Incentive Demonstration (HQID) project and the Premier Perspective[tm] data warehouse. We thank CMS for approving our use of the HQID data. Finally, thanks to Michelle Casey and Patricia Bland for their help with the Hospital Compare dataset.

Support for this paper was provided by the Office of Rural Health Policy, Health Resources and Services Administration, PHS Grant No. 5U1CRH03717-02-00.

Table of Contents

EXECUTIVE SUMMARY

The Premier Hospital Quality Incentive Demonstration (HQID) is a project sponsored by the Centers for Medicare and Medicaid Services (CMS) that is being conducted with participating hospitals that are members of Premier Inc., a national alliance of non-profit hospitals. The demonstration was designed to examine whether a system that explicitly pays-for-performance can lead to system-wide improvement in the quality of care that hospitals provide in selected medical conditions.

For this pay-for-performance demonstration project, 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 these medical conditions, statistical theory suggests that smaller hospitals can expect to experience much greater sampling variability in their performance scores.

In this study we construct statistical models to assess whether hospital size will impact our ability to identify "true" hospital ranks using data generated by pay-for-performance programs patterned after HQID. We address the following questions:

- How accurately can we expect to identify "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 relative performance is measured with substantial variation in accuracy?

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 for three medical conditions – acute myocardial infarction (AMI), heart failure (HF), and community acquired pneumonia (PN).

We obtained approval from Premier Inc. and from CMS to use the HQID data for Year 1. Our specific goal was to assess the implications of hospital size for the amount of uncertainty likely to exist in ranks of hospital annual composite scores. This study is not intended to be an evaluation of performance improvement brought about by the HQID.

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 three CAHs, and 44 rural hospitals. Currently CAHs constitute 23 percent of all short-term, general hospitals. Since the goal of the project is to specifically show the influence of small hospital size per se on the likely variability in hospital ranks, we conducted a second set of model runs using additional data gathered from the CMS Hospital Compare Program.

The results for community acquired pneumonia (PN) indicate that there is a dramatic inverse relationship between the size of the hospital and its expected range of ranking

positions for its "true" or stabilized mean rank. We conclude that the smallest hospitals among the augmented dataset would likely experience six times more uncertainty concerning their "true" ranks. Specifically, the smallest hospitals with 20 or less PN patients per year would have an average 95 percent Confidence Interval (CI) for their percentile-rank that spans 64 percentile points. For the largest hospitals with more than 1,100 PN patients per year the average 95 percent CI for their percentile-rank would span only ten percentile points.

With heart failure (HF) there also is an inverse relationship between the size of the hospital and its expected range of ranking positions for its "true" or stabilized mean rank. The smallest size stratum for HF among the augmented dataset (≤ 20 patients per year) would, on average, experience five times more uncertainty than the largest hospitals concerning their true relative performance.

For acute myocardial infarction (AMI), for hospitals with 20 or fewer AMI patients per year, the average width of the 95 percent CI for percentile-ranks is 63 percentile points, compared to average width of the 95 percent CI for percentile-ranks of nine percentile points for the very largest hospitals. The smallest hospitals would have seven times more uncertainty compared to the largest ones concerning their "true" rank.

In summary, we conclude that:

- All estimates of rank/percentile need to include adequate measures of uncertainty of those estimates.

- Identifying relative quality from single-year ranks based on composite scores will impact smaller institutions more severely 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 on the CMS website.

- It is reasonable to expect that increased uncertainty about a hospital's "true" relative quality level—arising from considerable annual variability in achieving the placement required for rewards and recognition—could impact their motivation to achieve higher quality levels.

- The likelihood and consequences of high levels of uncertainty concerning hospitals' relative levels of quality differs by specific medical condition, but in all cases they would have important implications for policy.

- The above findings are likely to generalize to hospitals beyond this sample.

- The results may under-state the degree of uncertainty like to be found in more mature P4P programs using simple ranks like the HQID. After several years of operation, the distribution of composite scores for all conditions would be expected to more closely cluster together, making ranks even more volatile.

- Using Bayesian models, we have begun to address the complex but policy-relevant issue of identifying and estimating the likely amount of uncertainty inherent in measuring relative quality through the ranks of composite scores. Assessing the likely impact of this uncertainty for P4P programs, and identifying ways of minimizing its effects, will be critical to their success.

INTRODUCTION

The Premier Hospital Quality Incentive Demonstration (HQID), sponsored by the Centers for Medicare and Medicaid Services (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 system-wide 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 pay-for-performance 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 relative performance is measured with substantial variation in accuracy?

This paper is organized as follows. First, we describe important features of the HQID project. Second, we summarize studies in the literature concerned with ranking. Third, we describe statistical models for analyzing the reliability of inferences based on ranking of hospitals' composite scores and also discuss the data used and their limitations. Fourth, we summarize our empirical findings. Finally, we explore policy implications that arise from including small, rural hospitals in pay-for-performance programs that use ranking to identify quality.

CMS/PREMIER HOSPITAL QUALITY INCENTIVE DEMONSTRATION

The Premier Hospital Quality Incentive Demonstration is a three-year demonstration project with data collected from participating hospitals from October 1, 2003 through September 30, 2004 (Year 1), October 1, 2004 through September 30, 2005 (Year 2), and October 1, 2005 through September 30, 2006 (Year 3). The HQID measures quality in five clinical areas: (1) Acute Myocardial Infarction (AMI), (2) Coronary Artery Bypass Graft (CABG) procedures, (3) Heart Failure (HF), (4) Community Acquired Pneumonia (PN) and (5) Hip and Knee Replacement (Hip/Knee) procedures. However, because

CABGs and Hip and Knee 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 three 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 Community Acquired Pneumonia (PN) as an example, a hospital receives a PN composite condition score based on the number of patients with PN that the hospital treated during the year needing these services who in fact received the services.

Table 2 shows the results for a specific hospital of the number of their patients "needing" and "receiving" the services of the seven PN measures. We use the term the "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. For this hospital, the composite quality score for their PN patients would be:

$$\text{Composite quality score} = \frac{\Sigma \,\#\,\text{Patients Received Service}}{\Sigma \,\#\,\text{Patients Needed Service}} = \frac{395}{457} = 86.4\%$$

By the nature of these quality measures (i.e. most measures are not applicable to all patients) varying proportions of patients with a given condition will be deemed to need the services related to a 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 in Years 1, 2 and 3

The rewards in the HQID are of two types:

- Financial incentives (increased payments), and
- 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 (N) varying by condition. The composite scores for the hospitals are then ranked, and from these ranks each hospital's percentile is determined [i.e. percentile = rank # / (N)].

A hospital with a composite clinical quality score in Year 1 that places it in the top decile—the 10^{th} percentile or higher of all hospitals in Year 1—receives a bonus of two

Table 1
Quality Measures Used in HQID

Clinical Conditions	Measures
Acute Myocardial Infarction (AMI)	1. Aspirin at arrival [1,2,3,4] 2. Aspirin prescribed at discharge [1,2,3,4] 3. ACEI for LVSD [1,2,3,4] 4. Smoking cessation advice/counseling [1,2,3] 5. Beta blocker prescribed at discharge [1,2,3,4] 6. Beta blocker at arrival [1,2,3,4] 7. Thrombolytic received within 30 minutes of hospital arrival [1,2,7] 8. PCI received within 120 minutes of hospital arrival [1,5,7] 9. Inpatient mortality rate [1,3,6,O]
Heart Failure (HF)	1. Left ventricular function (LVF) assessment [1,2,3,4] 2. Detailed discharge instructions [1,2,3] 3. ACEI for LVSD [1,2,3,4] 4. Smoking cessation advice/counseling [1,2,3]
Community Acquired Pneumonia (PN)	1. Percentage of patients who received an oxygenation assessment within 24 hours prior to or after hospital arrival [1,2,3,4] 2. Pneumococcal screening/vaccination [1,2,3,4] 3. Blood culture collected prior to first antibiotic administration [1,2,3] 4. Smoking cessation advice/counseling [1,2,3] 5. Initial antibiotic consistent with current recommendations [1,2,7] 6. Influenza screening/vaccination [1,2,7] 7. Antibiotic timing – percentage of pneumonia patients who received first dose of antibiotics within four hours after hospital arrival [1,2,4,7]

[1]National Quality Forum measure
[2]CMS 7th Scope of Work measure
[3]JCAHO Core Measure
[4]The National Voluntary Hospital Reporting Initiative (AHA Initiative)
[5]The Leapfrog Group proposed measure
[6]Risk adjusted using JCAHO methodology
[7]CMS and/or JCAHO to align with this measure in 2004
[O]Outcome measure (all other measures on this table are process measures)

Table 2
Example of Composite Clinical Condition Score Calculation

Community Acquired Pneumonia Measure	# Patients Needing Service	# Patients Receiving Service	% Receiving
Oxygenation assessment	92	92	100.0
Blood culture	80	73	91.3
Pneumococcal screening and/or vaccination	53	36	67.9
Antibiotic timing within 4 hours of arrival	91	74	81.3
Adult smoking cessation counseling	24	24	100.0
Flu screening/vaccination	35	20	57.1
Initial Antibiotic consistent with current recommendations	82	76	92.7
Total	457	395	86.4

percent of the Diagnosis Related Group (DRG)-based prospective payment for the patients with the condition among all Medicare fee-for-service (FFS) beneficiaries in Year 1 (CMS, 2005d). Hospitals with composite scores that place them in the second decile receive one 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 cut-off composite score for that clinical condition in Year 1 will have two 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 cut-off composite score but above the 90th percentile cut-off composite score from Year 1 will have one percent deducted from their Medicare payment for that condition in Year 3. (CMS, 2005a).

There is an important difference in the basis for determining rewards and penalties. Rewards are determined by the participating hospitals' rank each year. Penalties are determined not by the participating hospitals' rank in Year 3, but by their absolute performance (i.e. their composite clinical condition score) in Year 3. If all hospitals in Year 3 had composite clinical condition scores higher than the 80th percentile cut-off composite score in Year 1, none would be penalized. Rewards are based on relative performance – 20 percent of the hospitals will always be rewarded every year.

Finally, there is a public announcement on the CMS website that identifies hospitals that have placed in the top 50 percent of that year's ranking for each clinical condition, with these individual hospitals listed alphabetically. Hospitals that had composite scores

below the 50th percentile (the median) will not be named/identified on the CMS website (CMS, 2005d). This public announcement can be seen as both reward and penalty. Many hospitals may consider it the most important reward while many hospitals may consider the absence of their name from the CMS website to be the most important penalty. This could be an important public relations consideration for hospitals not in the top half of the rankings.

VARIABILITY IN COMPOSITE SCORES

We provide a brief explanation of why we expect to see 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 report we make reference to the expression "true" hospital rank. We mean by this the rank the hospital would have achieved in the "steady-state" if we could repeat the experiment of conducting the first year of the Premier/CMS HQID many times. Since this concept is critical for understanding the policy significance of this study's findings, we expand this discussion.

As noted, the composite scores are ranked by specific medical condition and hospitals are rewarded or penalized in various ways based on their scores for each year. One might ask: "that score and that rank is what actually happened in the hospital that year, so why shouldn't it be judged solely on the basis of it?" The fallacy of this position is best appreciated by the following counterfactual: Instead of having quality scores and ranks and hence rewards and penalties based on annual data, imagine the HQID were allowed instead to be based on weekly data (i.e. composite scores, ranks and rewards would all be determined weekly). No one would accept this because there would be too much variation from week to week in hospital composite scores and hence ranks and rewards and penalties. The ensuing amount of variability might well be expected to have significant, negative consequences for the motivation of participating hospitals and for achieving the goals of the P4P program. In any P4P program, including HQID, there will be some large hospitals that will have more patients with a specific condition in one week than some small hospitals will have in an entire year.

The critical, policy-relevant question suggested by this counterfactual is: *How much variability in composite quality scores and ranks due to sampling error is too much?* This study involves conceiving of a hospital's composite quality score as an estimate of its "steady-state" score. Consequently, this study also 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.

RELEVANT LITERATURE ON RANKING AND A BAYESIAN MODELING STRATEGY FOR HOSPITAL RANKINGS

There are 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, Louis et al., 2003), (Lin, Louis et al., 2004), (Liu et al. 2003), (Marshall and Spiegelhalter, 1998), (Andersson et al. 1998), (Goldstein and Spiegelhalter, 1996), (Howley and Gibberd, 2003), (Normand et al. 1997). We provide the conclusion and summary of just one:

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

> "It is clear that any attempt at using ranks either to compare clinics or summarize change over time may be seriously misleading even when, in this example, there are substantial differences between institutions. This is only to be expected as most institutions have overlapping intervals and hence precision in ranking is rarely obtainable, particularly for small institutions."

Of importance, they 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." For further information on Bayesian data analysis and examples of the use of Bayesian methods in other applications, see Carlin and Louis (2000), Gilks, Richardson and Spieglehalter

(1998), or Congdon (2003). Details of our Bayesian modeling strategy are presented in Appendix 1.

DATA SOURCES

We obtained approval from Premier Inc. and CMS to use the HQID data for Year 1. Our specific goal was to assess the implications of hospital size for the amount of uncertainty likely to exist in ranks of hospital annual composite scores. This study is not intended to be an evaluation of performance improvement brought about by the HQID.

We used Year 1 data on individual hospital-level reporting items from the participating hospitals. 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 beds and staffed beds, however, because the data on beds were incomplete, we present all empirical results by our size metric: hospitals' Net Number of Cases (NNC) for each medical condition.

For the AMI condition, some hospitals reported having no patients needing one or more of the eight measures. This poses no problem for calculating composite scores but it does introduce considerable complexity in our Bayesian models. As a consequence, we were able to model some hospitals having no patients needing one or more of the eight measures, but not all such hospitals. Our AMI model includes 243 hospitals out of 262 with Premier Inc. data.

Due to the same complexity issue raised for AMI, we use data from 263 of the 265 Premier hospitals in the PN model. We use 265 Premier hospitals for the HF model.

CMS 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 three 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 gathered from the CMS Hospital Compare Program for voluntarily participating hospitals (see Casey and Moscovice, 2005, for details). 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 the Hospital Compare participants that are CAHs.[1] The proportion of CAHs in the total augmented dataset (i.e. Premier Inc. hospitals plus Hospital Compare CAHs) was set equal to the proportion of CAHs (i.e. 23%) in the population of short-term, general hospitals.

There are some complications in using the Hospital Compare data. First, Hospital Compare data include all of the eight AMI process measures but neither AMI deaths nor the expected AMI mortality rate. Thus, the Standardized Survival Ratio cannot 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 purposes of the study. Thus, we use and report the results of the AMI model from the augmented dataset.

Secondly, there are differences in the duration of the collection periods for the Hospital Compare data. The Hospital Compare data had a starter set of measures that were later augmented by additional measures. Measures that were reported for less than or equal to nine months were normalized to represent full year measures.

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 3 provides composite scores associated with hospitals at the 10^{th}, 25^{th}, median, 75^{th} and 90^{th} 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 25^{th} to 75^{th} percentiles) have composite scores falling within a range of 11 composite score points (i.e. between 71 and 82). And 80 percent of the hospitals (from 10^{th} to 90^{th} percentiles) have composite scores within a range slightly less than 20 composite score points (i.e. 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 76.6 for PN.

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

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 nine percent of the hospitals are within a band of ± 2 composite points around the median value for HF, 69.6.

Lastly, AMI composite scores, based on 262 hospitals[2], exhibit the greatest amount of clustering. For AMI it takes only a range of 8.0 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).

Table 3
Distribution of Composite Scores for PN, HF and AMI

Percentile	Composite Scores		
	PN (%)	HF (%)	AMI (%)
10th percentile score	66.4	51.1	79.4
25th percentile score	71.0	59.8	85.6
Median	76.4	69.6	89.9
75th percentile score	82.1	80.2	93.5
90th percentile score	86.0	86.1	95.7
Lowest score	57.0	25.4	49.0
Highest score	92.4	96.4	99.4
Percentage of hospitals with composite scores ± 2 composite points around the median	23.0	9.0	29.0

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% credible interval (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% CIs of the hospitals in each stratum (i.e. the average number of ranking positions between the upper 95% CI rank value and the lower 95% CI rank value). Finally, we also provide the translation of ranking positions into the equivalent range of

[2] Note that we use the full 262 count of hospitals with AMI data for this exercise, not the smaller number feasible for modeling.

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.

Given the large amount of output generated by the models using both of the study samples, in this paper we provide and discuss only the model results derived from using the augmented sample of HQID participating hospitals plus the CAHs obtained from the Hospital Care program. There are two justifications for doing so: (i) this augmented sample with its additional set of smaller hospitals more completely illustrates the relationship between uncertainty and hospital size; and (ii) for the hospitals larger than the smallest ones that this augmented sample introduces, the implications for uncertainty are the same in the two samples. Full results of the modeling project are available from the authors.

Table 4 illustrates, for PN patients, the dramatic inverse relationship between the size of the hospital and its expected range of ranking positions about its "true" or stabilized mean rank. For the smallest hospitals—with 20 or less PN patients per year—the average range of ranking positions is 221 out of the 344 hospitals in this sample, or a full 64 percentile points. We emphasize that this is the average width of the 95 percent confidence intervals for the ranking positions of these smallest hospitals. 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, 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 Heart Failure there is also a strong inverse relationship between the size of the hospital and its expected range of ranking positions for its "true" or stabilized mean ranks (Table 5). 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 seven 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.

Table 4
Impact of the Number of Hospital Pneumonia Patients per Year on the Width of the 95% Confidence Intervals for Hospital Ranks and Percentile Values: Premier Inc. Hospitals Plus CAH Sample (n=344)

Annual Number of PN Patients in Hospital	Average Range of Rank Positions Falling within 95% CI for Ranks	Average Range of Percentile-Points Falling within 95% 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%
1001 – 1100	38	11%
1101 – 2313	35	10%

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 scores for the two conditions as provided in Table 3, since the distribution of HF composite scores was spread out much more than was the case for PN.[3] A less concentrated distribution of composite scores for HF is the equivalent of a stronger signal, or more information about true relative performance.

[3] The distribution of composite scores in Table 3 is for only the Premier Inc. hospitals.

Table 5
Impact of the Number of Hospital Heart Failure Patients per Year on the Width of the 95% Confidence Intervals for Hospital Ranks and Percentile Values: Premier Inc. Hospitals Plus CAH Sample (n=348)

Annual Number of HF Patients in Hospital	Average Range of Rank Positions Falling within 95% CI for Ranks	Average Range of Percentile-Points Falling within 95% 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%
1001 – 1100	28	8%
1101 – 1926	25	7%

For AMI (Table 6), we would expect to see the greatest amount of uncertainty displayed in true ranks based on the results of Table 3, and for the most part we do. For 20 or fewer AMI patients per year, the average width of the 95% 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.

Table 6
Impact of the Number of Hospital AMI Patients per Year on the Width of the 95% Confidence Intervals for Hospital Ranks and Percentile Values: Premier Inc. Hospitals Plus CAH Sample (n=314)

Annual Number of AMI Patients in Hospital	Average Range of Rank Positions Falling within 95% CI for Ranks	Average Range of Percentile-Points Falling within 95% 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 – 1000	40	13%
1001 – 1100	43	14%
1101 – 1926	29	9%

Uncertainty about "True" Placement of the Top 20 Percent of Hospital Ranks

We summarize the uncertainty of hospital placement in the top 20 percent in Table 7 since this is the specific way that ranks are used in the HQID for assigning rewards.

Table 7 contains the following measures of uncertainty:

- Hospitals are ranked by their Bayesian model probabilities of being in the top two deciles (20 percentile or better). Among the hospitals within the top 20 percent of this ranking, we report the proportion having Bayesian model probabilities of being in the top 20 percent 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. In an ideal world we would like all hospitals assigned to the top 20 percent to have very high probabilities of having a true rank high enough to place them in the top 20 percent. Since we are dealing with estimates of relative performance based on relatively small numbers of patients, we have to accept the reality of something less than the ideal high probabilities of being truly in the top 20 percent. 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 identified in the last step above, the average probability of being in the top 20 percent. This measure reflects how quickly or slowly 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.

Table 7
Measures of Uncertainty Concerning Top 20 Percent Placement in Rank: Premier Inc. Hospital Plus CAH Sample

	PN	HF	AMI
Percentage of hospitals placed in top 20% that have 95% or greater probability of being in top 20%	52%	57%	49%
Percentage of hospitals placed in top 20% that have less than 80% probability of being in top 20%	33%	23%	37%
Average probability of being in top 20% for the hospitals that have less than 80% probability of being in top 20%	61%	67%	57%
Probability of being in top 20% for the last hospital assigned to the top 20%	43%	55%	43%

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% 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 takeaway 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% that would bring public recognition on the CMS website.

- It is reasonable to expect that increased uncertainty about a hospital's "true" relative quality level—arising from considerable annual variability in achieving the placement required for rewards and recognition—could impact their motivation to achieve higher quality levels.

- 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-important way. From the literature cited, we have a high expectation of encountering difficulty in gauging true relative performance from ranks based on annual composite scores. Moreover, the size of the group of hospitals being compared should not alter the generalizability of these results.

- The results may under-state the degree of uncertainty likely to be found in more mature P4P programs using simple ranks 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.

- Using Bayesian models, we have begun to address the complex but policy-relevant issue of identifying and estimating the likely amount of uncertainty inherent in measuring relative quality through the ranks of composite scores. Assessing the likely impact of this uncertainty for P4P programs, and identifying ways of minimizing its effects, will be critical to their success.

REFERENCES

Andersson, J., Carling, K., and Mattson, S. "Random Ranking of Hospitals is Unsound." *Chance* 11:34-39, 1998.

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

Casey, M. and Moscovice, I. "CAH Participation and Initial Results of Hospital Compare Data." *Flex Monitoring Team Briefing Paper No. 8*. Minneapolis, MN: Upper Midwest Rural Health Research Center, University of Minnesota, 2005.

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

Centers for Medicare and Medicaid Services (CMS). Premier Hospital Quality Incentive Demonstration, *Fact Sheet*. Available at: http://www.cms.hhs.gov/researchers/demos/phqi/default.asp, 2005a.

Centers for Medicare and Medicaid Services (CMS). Premier Hospital Quality Incentive Demonstration, *Historical Data FAQs.* Available at: http://www.cms.hhs.gov/quality/hospital/HistoricalDataFAQ.pdf, 2005b.

Centers for Medicare and Medicaid Services (CMS). Premier Hospital Quality Incentive Demonstration, *Clinical Areas & Quality Measures*. Available at: http://www.cms.hhs.gov/quality/hospital/PremierMeasures.pdf, 2005c.

Centers for Medicare and Medicaid Services (CMS). Premier Hospital Quality Incentive Demonstration, *Terms and Conditions.* Available at: http://www.cms.hhs.gov/HospitalQualityInits/downloads/HospitalTermsAndConditions20060601.pdf, 2005d.

Gilks, W., Richardson, S., and Spiegelhalter, D. *Markov Chain Monte Carlo in Practice.* Boca Raton FL: Chapman and Hall/CRC Press, 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* 159:385-443, 1996.

Howley, P. and Gibberd, R. "Using Hierarchical Models to Analyze Clinical Indicators: A Comparison of the Gamma-Poisson and Beta-Binomial Models." *International Journal for Quality in Health Care* 15:319-329, 2003.

Lin, R., Louis, T., Paddock, S., and Ridgeway, G. "Loss Function Based Ranking in Two-Stage, Hierarchical Models." Johns Hopkins Department of Biostatistics Working Paper 6. Available at: http://www.bepress.com/jhubiostat/paper6, 2003.

Lin, R., Louis, T., Paddock, S., and Ridgeway, G. "Ranking USRDS Provider-Specific SMRs from 1998-2001." Johns Hopkins Department of Biostatistics Working Paper 67. Available at: http://www.bepress.com/jhubiostat/paper67, 2004

Liu, J., Thomas A., Pan, W., Ma, J., and Collins, A. "Methods for Estimating and Interpreting Provider-Specific Standardized Mortality Ratios." *Health Services & Outcomes Research Methodology* 4:135-149, 2003.

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

Louis, T. and Shen, W. "Innovations in Bayes and Empirical Bayes Methods: Estimating Parameters, Populations and Ranks." *Statistics in Medicine* 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* 316:1701-1705, 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* 92:803-814. 1997.

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

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. *WinBUGS User Manual*, Version 1.4, January 2003.

Appendix 1

Details of a Bayesian Modeling Strategy for Hospital Rankings

For the two conditions of heart failure (HF) and community-acquired pneumonia (PN), and for the eight process measures available for acute myocardial infarction (AMI), our Bayesian models all share a common structure.[4] Using HF as an example, for each of the four HF process measures in 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 in that hospital in that year with the diagnosis of HF that were not transfers into or out of the hospital. We then posit that for each hospital [i from 1 to 265], the number of cases 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].

We use Bayesian models to obtain each hospital's "true" or "steady-state" values of these two sets of binomial parameters: the proportion of the Net Cases needing each measure's service and 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 don't 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-time intensive sampling iterations, Bayesian models allow us to derive estimates of the values of the parameters from each of the modeling levels. Once we are convinced that 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, through sensitivity analyses, 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 that 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 8 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.

[4] We discuss the standardized survival ratio, the outcome measure for AMI below.

- 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 iteration of the post-convergence simulation we use the individual-hospital composite scores 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 estimate of the "true" or "steady-state" value of its rank.

- We also obtain the full posterior distribution of each hospital's ranks and from this we directly assess the range of each hospital's ranks. In particular, we can readily show the 95% confidence intervals for the mean ranks—or 95% credible intervals in Bayesian modeling. These 95% 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, we derive a second ranking estimator. 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 9^{th} outcome measure, the standardized survival ratio (SSR). We use a separate Bayesian hierarchical model 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 to 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% of the composite score from the eight process measures plus 11% of its SSR value.