MATHEMATICA
Policy Research

FINAL REPORT

Improving the AHRQ Quality Indicators: Summary of Findings and Recommendations for Improving the Methodological Approach

December 22, 2014

David Jones
Eric Schone
Frank Yoon
Alex Bohl
Sheng Wang
Mariel Finucane

Submitted to:
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

Project Officer: Pamela Owens
Contract Number: HHSA2902012000041 Task Order No. 2

Submitted by:
Mathematica Policy Research
955 Massachusetts Avenue
Suite 801
Cambridge, MA 02139
Telephone: (617) 491-7900
Facsimile: (617) 491-8044

Project Director: David Jones
Reference Number: 40156

This page has been left blank for double-sided copying.

CONTENTS

TABLES

This page has been left blank for double-sided copying.

FIGURES

This page has been left blank for double-sided copying.

EXECUTIVE SUMMARY

The Agency for Healthcare Research and Quality (AHRQ) Quality Indicators™ (QIs) were developed to help states assess inpatient quality of care at hospitals. AHRQ developed three categories (modules) of QIs that estimate rates of different types of adverse events at the hospital level: Patient Safety Indicators (PSIs), Inpatient Quality Indicators (IQIs), and Pediatric Safety Indicators (PDIs). The QIs were soon also used by hospitals to monitor their performance regarding patient safety and mortality. Furthermore, a demand for comparisons of quality between hospitals for various public and private programs led to the risk and reliability adjustment of the QIs. The leveling feature of the risk-adjustment process and the adjustment for the reliability of estimated hospital rates of adverse events facilitated use of the QIs to compare hospital quality in quality-improvement initiatives. Considering the high profile and high-stakes uses of the indicators in comparative reporting programs, AHRQ has made it a priority to identify threats to the validity of the QIs for use in hospital comparisons.

The suitability of QIs for use in comparing hospitals' quality depends on the efficacy of their risk and reliability adjustment. Because the comparisons are made between hospitals with different patients, the user must account for differences in the risk of an adverse event in the two patient populations. In addition, because the comparisons are made between hospitals with different amounts of available performance information, they must account for the reliability of the estimates of quality. The raw rates of adverse events estimated by the QIs are adjusted based on patient discharge records to account for factors that increase or decrease a patient's risk for a given adverse event but which are not influenced by the quality of care delivered to the patient (for example, a patient's gender, age, or comorbidities that are present at the time of admission). The risk-adjusted rate is calculated by indirect standardization; that is, a hospital's rate can be interpreted as the performance of a hospital treating its patients relative to a hypothetical average hospital treating patients with the same characteristics. The risk-adjusted rates are then reliability adjusted (shrunken) to account for uncertainty about a hospital's rate arising from the limited information about its performance contained in its discharge records. Through AHRQ's approach to reliability adjustment, the risk-adjusted rates are the weighted average of the hospital's own rate and a reference population rate believed to provide an estimate of the hospital's likely performance in the absence of any information from its own discharges.

To avoid mischaracterization of hospital quality and produce the comparisons that designers of programs using the QIs intended, AHRQ aims to ensure that the leveling produced by risk adjustment is fair and accurate and that the reliability adjustment produces the most accurate estimates possible given the available information in the discharge records. Stakeholders and researchers have observed systematic variations by hospital type in estimates of QI rates. These variations are a sign of a possible problem (that is, the differences could be caused by a factor other than quality, such as unmeasured risk), but also a potential avenue of improvement in the QIs that would improve the suitability for use in hospital comparisons.

The objective of this project is to make recommendations regarding modifications to AHRQ QI methods and suggest topics for related research. To achieve this objective, we studied the differences in the AHRQ QI rates across hospital types, reviewed methods used to estimate the rates, and tested modifications to the methods. In particular, we focused on modifications to the risk-adjustment and reliability-adjustment methods.[1] Through our review of risk- and reliability-adjustment methods, we identified four specific areas in which opportunities for improvement could be found. We tested modifications in each of the areas that could lead to greater accuracy in hospital comparisons using the QI rates. The four areas are: the method used to standardize hospital rates (indirect versus direct standardization), incorporation of hospital characteristics in risk adjustment, shrinking (also referred to as smoothing) reliability-adjusted rates to targets that vary according to hospital type or the characteristics of the study sample, and implementing a formal empirical Bayes or Bayesian statistical framework for estimating reliability adjusted rates using alternate underlying assumptions regarding the distributions of hospital rates (Bohl et al. 2014; Chen et al. 2014; Jones et al. 2014a, b; Wang et al. 2014). We summarize the methodological challenges targeted for improvement and the potential improvements examined in Table ES.1.

We tested these modifications and used the findings in the analyses to support recommendations for potential modifications and to identify areas requiring additional study. In this report, we summarize the findings from the analyses, highlight recommendations that can be gleaned from the analyses, discuss considerations based on strengths and weaknesses of the current and modified methods, and logical extensions of the analyses.

[1] Assessing modifications to the discharge- or patient-level variables included in the risk-adjustment models, the overall methodological framework for risk and reliability adjustment, and compositing methods are outside of the scope of this project. We discuss potential extensions of the analyses conducted under this project and logical next steps in subsequent sections.

Table ES.1. Summary of modifications tested

Area of analysis	Methodological challenge	Potential improvement
Risk adjustment		
Standardization approach	The current approach of estimating indirectly standardized rates might not adequately capture the effect of differences in case mix by hospital type.	We compare indirectly and directly standardized rates to assess the current approach and discuss possible instances in which direct standardization could be an improvement.
Risk-adjustment models	Differences in hospital rates by hospital type could reflect differences in unmeasured risk rather than differences in hospital quality.	We examine the inclusion of indicators for hospital type in the risk-adjustment models.
Reliability adjustment		
Shrinkage targets	The current shrinkage target (mean rate of the 44-state HCUP reference population) might not be appropriate for some populations. In addition, peer group mean rates could provide more appropriate shrinkage targets.	We examine alternate shrinkage targets for various populations as well as shrinking to peer group means.
Empirical Bayes framework	The current QI specifications do not clearly state the empirical Bayes formulation. In addition, the apparent assumption of a normal prior for hospital rates may be overly restrictive and misaligned with the data on hospital quality.	We test the implementation of a formal empirical Bayes framework and examine the effect of alternate distributional assumptions.

HCUP = Healthcare Cost and Utilization Project.

Data

To support the analyses, we constructed an analytic file of inpatient discharge records merged with hospital characteristics. The file contains discharges from the State Inpatient Databases (SID), obtained from the Healthcare Cost and Utilization Project (HCUP) Central Distributor, coordinated by AHRQ, which is the same source as the reference population used by AHRQ to estimate the risk-adjustment models. The SID contain inpatient discharge abstracts for patients of all ages and for all payers admitted to all community hospitals in participating states. The analytic file contains data obtained from 12 states in 2009 and 2010.[2]

We merged hospital-level information from several sources onto each discharge using hospital identifiers. Many of the hospital characteristics were obtained from the 2010 American Hospital Association Survey Database; other sources include the 2010 Centers for Medicare & Medicaid Services (CMS) Impact File, CMS Certification Numbers, 2013 United States

[2] We would like to acknowledge the HCUP Data Partners: Arizona Department of Health Services, Arkansas Department of Health, California Office of Statewide Health Planning and Development, Florida Agency for Health Care Administration, Iowa Hospital Association, Kentucky Cabinet for Health and Family Services, Maryland Health Services Cost Review Commission, Massachusetts Center for Health Information and Analysis, Nebraska Hospital Association, New Jersey Department of Health, New York State Department of Health, and Washington State Department of Health.

Department of Agriculture Economic Research Service Rural-Urban Continuum Codes, and aggregate information from the discharge data, such as the percentage of discharges for various primary payers (Medicaid, Medicare, and uncompensated care). For the purpose of organizing and presenting findings, we grouped the hospital characteristics into three categories: (1) structural characteristics, which include organizational and operational characteristics of hospitals related to levels of resources and experience (for example, teaching status and bed size); (2) aggregate patient characteristics, which represent factors that might indicate unmeasured individual patient risk or factors that directly affect hospital resources (for example, disproportionate share status [DSH]); and (3) market and local area characteristics, which represent factors external to the hospital that can affect the primary population served by hospitals, and thus unmeasured individual patient risk (for example, critical access hospitals [CAHs]).

We used the analytic file to create hospital-level results according to the current QI methods using the specifications in the AHRQ QI software without modification. Then, we modified the methods and tested the effects of the modifications by comparing the results by hospital type to those generated using the unmodified current methods.

We began by examining all risk-adjusted PSIs, IQIs, and PDIs and a large set of hospital characteristics in a literature review and an exploratory data analysis (EDA). For in-depth analyses of modifications to the methods, we focused on a subset of QIs and hospital characteristics identified from the literature review and EDA. We selected a subset of QIs to maximize the generalizability of the findings in this report to the full set of QIs. Hence, we chose QIs that represent a range of clinical properties (such as PSIs addressing continuity of care as well as technical care) and statistical properties (such as including QIs that measure rates of relatively rare events as well as more common events). In addition, because composite indicators are used prominently in federal programs, we prioritized QIs that have the largest weights in the calculations of hospital composite values. We selected the hospital characteristics that demonstrated an empirical relationship with QI results in the literature, have a strong conceptual rationale for such a relationship, and are important in the policy context of making hospital comparisons. Although the specific combinations of QIs and hospital characteristics varied slightly by analysis, the primary relationships examined are between the QIs and hospital characteristics listed in Table ES.2.

Table ES.2. AHRQ QIs and hospital characteristics included in the analysis

Quality indicators	Hospital characteristics
• PSI 06 Iatrogenic Pneumothorax Rate • PSI 12 Postoperative PE/DVT Rate • PSI 13 Postoperative Sepsis Rate • PSI 14 Postoperative Wound Dehiscence Rate • PSI15 Accidental Puncture or Laceration Rate • IQI 15 Acute Myocardial Infarction Mortality Rate • IQI 16 Heart Failure Mortality Rate • IQI 20 Pneumonia Mortality Rate • PDI 01 Accidental Puncture or Laceration Rate • PDI 10 Postoperative Sepsis Rate • PDI 12 Central Venous Catheter-Related BSI Rate	• Number of licensed hospital beds: broken into indicators for bed size quartiles • Teaching hospital status: indicator for major/minor teaching status • Disproportionate share (DSH) status: indicator for greater than 15 percent of patient populations composed of disproportionate share patients • Critical access hospital (CAH) status: indicator for designation as a critical access hospital

PE/DVT = pulmonary embolism or deep vein thrombosis; BSI = blood stream infection.

Literature review

There were no definitive relationships between the QIs and hospital characteristics (that is, no relationships were completely consistent in direction and statistical significance of the association) across studies examined in the literature review (Dy et al. 2013). Many of the studies found that the estimated associations between the QI rates and hospital characteristics studied were not statistically significant. However, for some combinations of the QIs and hospital characteristics, the associations were either in the same direction or not statistically significant. For example, for bed size and 8 of 15 PSIs examined, approximately half of the studies found that greater number of beds was associated with higher odds of patient safety events, whereas half of the studies found no statistically significant association; there were no statistically significant associations for bed size and the other 7 PSIs. Surprisingly, there was little mention of what factors might contribute to the observed associations in the studies reviewed. In addition, the review did not uncover any studies that directly assess the methodological approach AHRQ uses to estimate the QIs results.

Exploratory data analysis

In comparing relationships of QI rates and hospital characteristics, the most common pattern observed is that many hospital characteristics have one association with the patient safety-related QIs (PSIs and PDIs) but the opposite association with IQIs. Several structural and market/local area characteristics exhibited this pattern (bed size, teaching hospitals, high nurse staffing ratios, and, for PDIs, children's hospitals) as did two characteristics that describe the location or market of hospitals (non-CAHs and hospitals located in urban settings). In addition, hospital bed size was the most consistent (in terms of statistically significant associations with QI rates) and strongest (in terms of the magnitude of the associations) predictor of hospital rates in a multivariate analysis; whereas many of the other hospital characteristics (particularly those highly correlated with volume, such as teaching status and urban location) were not as strong.

We also observed that some hospital types had higher rates (lower quality) for indicators in all three modules. This pattern is particularly true for aggregate patient characteristics, such as the hospital's DSH status, its proportion of Medicaid discharges, and its proportion of Medicare discharges. The relationships are particularly consistent and strong for DSH status, which also exhibited consistently higher rates in the multivariate regression analysis.

The EDA did not uncover evidence to support or refute the hypotheses that differences in coding practices, differences in risk, or the volume–QI relationship account for the differences observed in hospital QI rates by hospital type.

Analysis of modifications to the QI methods

We examined two modifications to the risk-adjustment methods: an alternate approach to the standardization of rates and the incorporation of hospital characteristics in the risk risk-adjustment models. We also examined two modifications to the reliability-adjustment methods: shrinking to alternate shrinkage targets and implementing a formal empirical Bayes or Bayesian framework with alternate assumptions regarding the distribution of hospitals' rates.

1. Assessing benefits and limitations of indirect standardization

We studied the effect of standardization methods by comparing the differences in risk-adjusted rates between hospital types when the rates are directly standardized with differences in rates when indirectly standardized. By direct standardization, we compare exactly the same types of patients based on their risk factors between hospital types. This approach enables us to compare performance between hospital types measured by QI rates on their care for the same case mix of patients.

We compared whether the differences by hospital type were statistically significant for the rates calculated by indirect and direct standardization. We calculated two 95 percent confidence intervals (CIs) for differences in risk-adjusted rates between hospital types: one CI produced by direct standardization and the other by indirect standardization. If the 95 percent CI for the difference in rates between two hospital types included zero, then the two hospital types were not considered to be statistically different by the given standardization method. A change in the statistical significance of a difference by hospital type when we estimated rates by direct standardization instead of indirect is evidence that risk-adjustment may be improved, either through direct standardization or changes to model specifications to better support comparisons of QI rates across hospital types using indirect standardization. For this analysis we focused on the relationships between QIs (five IQIs, four PSIs, and 3 PDIs) and two hospital characteristics (teaching/nonteaching hospitals and the smallest/largest hospitals by total bed size quartiles).

To compare directly standardized rates between hospital types, we required combinations of risk factors to be present in discharges for both hospital types. If the combinations did not exist in both hospital types, it would suggest that the hospital types are not comparable. However, in general, patient characteristics were similar between hospital types before stratification. Thus, it was possible to create matched discharge groups across hospital types to create directly standardized rates based on similar patient populations; less than one half of one percent of discharges were dropped for the PSIs (with the exception of PSI 12) and IQIs because the risk profiles did not match between the hospital types. In contrast, for three QIs (PSI 12, PDI 10, and PDI 12), five to six percent of discharges could not be matched across hospital types, suggesting noteworthy differences in patient populations.

We found that for most combinations of QIs and hospital characteristics examined in the analysis, indirectly and directly standardized rates produced the same conclusion regarding differences in QI rates by hospital type. For 2 of the 10 comparisons made for the IQIs, the conclusion drawn about the differences in rates between hospital types changed when using directly standardized rates (IQI 11 and bed size; IQI 19 and teaching status). The rate of disagreement was even lower for the PSIs; the conclusion differed for 1 of the 8 PSIs (PSI 15 and bed size). The PDIs showed the highest rate of disagreement; the conclusion differed for 2 of 6 PDIs (PDIs 1 and 10 and teaching status).

Table ES.3. Recommendations for standardization approach

Category	Recommendations
Modification to methods	For most QIs, risk adjustment through indirect standardization adequately adjusts for different observed case mixes of patients between hospital types in comparison with direct standardization. In the cases for which this is not true, we recommend further analysis to consider alternative specifications of the models to improve comparisons using direct or indirect standardization.
Target audiences	For comparisons across hospitals with a range of characteristics, indirect standardization is an appropriate method. The approach establishes an appropriate benchmark against which each hospital is compared. For objectives other than national, population-level assessments of hospital performance, risk adjustment through direct standardization methods could offer benefits over indirect standardization. The approach could be appropriate for any user aiming to assess performance for a specific population rather than the average population nationwide. The AHRQ QI software could include a template containing risk-adjusted rates by risk profiles against which a user may compare a discharge sample from a collection of one or more hospitals.
Future analysis	More in-depth investigation of direct standardization approaches can identify specific QIs and populations for which that method is the best approach (for example, uses/users that would like to compare multiple hospitals' performance over a similar population). In addition, neither direct nor indirect standardization solves the problem that there could be differences in unmeasured risk factors that contribute to differences in QI rates by hospital types. Further analysis is needed to isolate differences in quality by hospital type apart from factors such as unmeasured risk.

2. Incorporating hospital characteristics in the risk-adjustment models

We tested whether incorporating an indicator of hospital type in the risk-adjustment models as an additional risk factor could lead to an improvement in the accuracy of hospital comparisons across hospital types. If the differences in risk-adjusted rates documented in the EDA are due to unknown or unmeasured differences in patients that are correlated with risk of adverse events, including hospital characteristics in the risk-adjustment model could help account for these factors and further level comparisons of hospitals with different patient populations. For example, if teaching hospitals have higher rates of patient safety events because they treat higher proportions of patients with a particular risk factor not identified in patient records, the difference in mean rates between the two hospital types would be attributed to that risk factor. Thus, the predicted values for their patients would be adjusted upward, leading to a higher expected rate of events at teaching hospitals and lower observed to expected ratios and risk-adjusted rates. Unfortunately, if differences in rates across hospitals reflect differences in quality across hospital types, these differences in quality would also be obscured by this adjustment, because average differences between hospital types incorporated in the models are adjusted to zero.

We find that the estimated relationships between hospital characteristics and adverse events when adding hospital type indictors to the risk-adjustment models were consistent with the findings of the EDA. Large hospitals, teaching hospitals, and non-CAHs have higher estimated risk-adjusted rates for the PSIs and PDIs, but lower rates for IQIs; whereas DSH have higher estimated rates for all three of the QI modules. The size and precision of the estimated relationships varies among the QIs, with some showing fairly strong relationships and others small and/or imprecisely measured relationships (for example, PSI 12 demonstrates a fairly large association even accounting for the discharge-level risk factor information), but there does not appear to be a pattern in the relationships by the clinical or statistical properties of the QIs. As expected, the effect on QI rates is to largely remove any variation in mean rates between hospital types. In addition, we find that the degree to which model performance improves depends on the specific QIs and hospital characteristics. When the measured associations are relatively large, the hospital characteristics tend to add to the models' fit. However, there is also evidence that none of the hospital characteristics add much to predictive power for any of the QIs, and adding hospital characteristics does little to change the magnitude and precision of the estimated relationships between the current discharge-level risk factors and outcomes.[3]

Taken together, the findings from the modified models indicate that hospital characteristics explain additional variation in adverse events for some combinations of the QIs and hospital characteristics above the variation currently explained by the discharge-level risk factors. If these relationships indicate differences in unmeasured risk between hospitals types, adding hospital type indicators could improve the accuracy of estimated rates and hospital comparisons. However, the origins of these relationships (role of risk versus quality), which will play a substantial role in determining whether the modification to the models will improve or reduce the accuracy of hospital comparisons, remain unclear.

In the absence of definitive evidence regarding the roles that differences in risk and quality play in driving the observed differences in hospital QI rates by hospital type, we assessed whether incorporating hospital characteristics in the risk-adjustment models improved the accuracy of hospital comparisons by simulating variations in the proportion of the differences due to risk and quality. We used information from the estimated relationships between hospital QI rates and hospital types from the models to simulate hospital discharge data for which the differences in rates are due to a range of risk and quality mixes that we defined (all risk, all

[3] We also tested two alternate approaches for incorporating hospital characteristics in the risk-adjustment models: (1) splitting the sample by hospital type and reestimating the models separately on the samples (in effect, a fully interactive model) and (2) adding an average hospital type effect in the calculation of predicted values, in which all hospitals receive an adjustment to their expected rate based on the average hospital type effect. The latter approach is intended to account for potential correlation of hospital characteristics with discharge-level risk factors included in the models (Ash et al. 2011). The resulting expected rates are the predicted rates of adverse events for the average hospital type and with an average patient case mix. We find that the hospital QI rates and rankings for the split models are nearly identical to those for the primary modified approach of adding hospital type indicators. We find that the hospital rankings for the model incorporating an average effect are nearly identical to the current or base models, which is not surprising because most of the estimated average effects are quite small and the size of the effect applied to predicted values is the same for all discharges and hospitals.

quality, and mixtures of both). We then tested how the modified models performed in ranking hospital quality compared to the current models.

In all but one instance, the inclusion of hospital characteristics as risk factors (we focused on teaching status and bed size) obscured the quality signal contained in the risk-adjusted rates. Risk-adjusted rates containing hospital type risk factors showed reduced correlation of the risk-adjusted rate with simulated "true" hospital quality compared to the base model. The exception to this finding occurred when the difference in hospital type mean effects was assumed to be entirely due to risk. Furthermore, although the addition of either teaching status or bed size to the models reduced the ability to rank all hospitals together according to simulated quality, it had little effect on the models' ability to rank hospitals within hospital types; that is, within hospital type, the correlations of risk-adjusted rates with simulated quality were nearly identical in models with and without the hospital type indicators.

Table ES.4. Recommendations for incorporating hospital characteristics in risk-adjustment models

Category	Recommendations
Modification to methods	The evidence from the simulation analysis suggests that including hospital type indicators as risk factors is not advisable for any use that compares QI rates across hospital types and when the user is not certain that nearly all of the observed differences in QI rates are due to differences in unmeasured risk.
Target audiences	If users of QI results determine that the potential differences in risk are important enough to be considered—but the ramifications of adjusting away differences in quality by hospital in making comparisons across all hospitals are too great—they might also consider a more restricted method of comparisons rather than altering the risk-adjustment methods. For example, providing comparisons to both a national benchmark and a peer group benchmark (stratification or peer grouping) could provide the information needed for quality improvement while ensuring that differences in hospital quality related to hospital type are not ignored.
Future analysis	We recommend examining the inclusion of discharge-level variables (such as measures of clinical, sociodemographic, or socioeconomic characteristics) that could proxy for patient risk in the risk-adjustment models as a potential solution to hypothesized differences in unmeasured risk by hospital type. Adding discharge-level variables as proxies for risk allows for potential improvement (accounting for unmeasured risk that varies by hospital type on average) without necessarily obscuring differences in quality by hospital type (as long as patients with the given discharge-level factor do not receive lower quality of care on average). However, there is still the same danger of obscuring quality if the discharge-level variable is correlated with quality (indication of quality delivered to hospitals where these patients receive care) rather than directly influencing quality or performance (a risk factor). Thus, we also recommend further analysis to provide evidence regarding the factors driving the differences in QI rates associated with these patient characteristics.

3. Shrinking to alternate targets

We tested two changes to the method of reliability adjustment, which shrinks hospitals rates to a single reference population rate using reliability weights estimated over the reference population: (1) shrinkage targets based on the sample of discharges and hospitals being analyzed by the user (for example Medicare patients only) and (2) separate shrinkage targets for different hospital types (for example, separate shrinkage targets for small and large hospitals). The objective is to estimate shrinkage targets that are better representations of true hospital quality for the population to be analyzed than provided currently using the reference population, leading to more valid comparisons of quality across hospital types. When comparing the alternative approaches to reliability-adjustment, we reviewed their effects overall and by hospital type on: (1) shrinkage targets (mean hospital rates), (2) average reliability weights, and (3) hospital's adjusted rates and the performance categories assigned by comparing hospitals' reliability-adjusted rates to the national mean rate; hospitals are classified as better than, no different from, or worse than the national mean rate. We examined the changes to these results for two types of analytic samples, the 12-state 2010 SID, which is a subsample of the HCUP reference population, and all Medicare fee-for-service discharges at inpatient prospective payment system (IPPS) hospitals from April 2011 through March 2012, which represents a sample external to the reference population.

The overarching finding is that although the magnitude of the effect of the alternate approaches on mean hospital QI rates and reliability varies depending on the QI and the specific approach, the effects on comparative performance (either classifying hospitals into performance categories or ranking hospitals) are typically small. The largest observed change to hospital classification is a shift in the performance classification for roughly eight percent of hospitals for one of the QIs, but the percentage is below one percent in the majority of cases, and the hospital rankings are nearly identical between the base and modified models. In addition, the findings are similar when examining the 12-state sample of SID and the Medicare population.

Small hospital reliability-adjusted rates are the most sensitive to a change in the shrinkage target, which is not surprising given that, on average, they are pulled to the target to a greater extent. However, on the whole, there is little movement in performance categorization. This finding is largely due to the large confidence intervals on average for QI rates and the restrictiveness of the method for classification. The largest movement in performance categorization is for large hospitals, which are more likely to be in either outlier category (better than or worse than the national mean rate) under the current AHRQ approach or the modified approaches because, on average, they have narrower confidence intervals, increasing the likelihood that the interval overlaps with the national mean rate.

Table ES.5. Recommendations for shrinking to alternate targets

Category	Recommendations
Modification to methods	There is a strong conceptual rationale for implementing shrinkage targets estimated on the analytic sample when feasible, and there are potentially substantial effects on QI rates of doing so, depending on the QI. However, further consideration is needed to determine if the approach is appropriate for different uses and users of the QIs.
Target audiences	Reestimating shrinkage targets could be an appealing approach for users with populations containing many hospitals that are substantially different than the reference population. Furthermore, estimating peer group shrinkage targets based on hospital type could be a promising approach if differences in estimated rates by hospital type represent differences in true quality. Large numbers of discharges are needed for users wishing to reestimate the models. In addition, the complexity of reestimation is such that most users will not be able to reestimate without assistance. Thus, if the modifications to the approach are determined to be a valuable option for users of the QIs, implementing the modifications may require AHRQ to add the necessary flexibility to the QI software to reestimate some or all of the parameters needed to estimate hospitals' rates (while maintaining the ability of individual hospitals to generate their rates).
Future analysis	We recommend further analysis to consider whether shrinking to targets estimated within sample is appropriate for various users of the QIs. Regarding shrinking to peer group targets, the approach has the potential to be restrictive, and AHRQ must be confident that differences by hospitals type reflect true differences in quality. Otherwise, the rates for hospitals of a given type will be partially (and potentially artificially) restricted to their group mean, and differences between that mean and the mean of their counterparts could reflect something other than quality (for example, differences in unmeasured patient risk by hospital type). We also recommend further analysis on the potential of shrinking risk-adjusted rates using other information, such as a hospital's rate in a previous year or using information from other QIs for a given hospital to reliability adjust a rate for a different QI.

4. Implementing an empirical Bayes or Bayesian framework

We implemented a formal empirical Bayes or Bayesian framework and examined the effect of alternate prior distributions that may improve the accuracy of estimated reliability-adjusted rates. We investigated prior distributions that could provide a better fit to the hospital discharge data than the current assumption of a normal distribution. For example, the EDA demonstrates that the AHRQ QI risk-adjusted rates are correlated with hospital characteristics, such as teaching hospital status and bed size (Jones et al. 2014a). Additionally, the rates for some QIs exhibit skewed distributions with mass points at zero and heavy tails, suggesting the existence of outliers. In addition, we investigated the implications of implementing a formal empirical Bayes framework with the explicit assumption that hospital rates are normally distributed to the current approach which uses similar principles and an implicit assumption that rates are normally distributed.

We calculated hospital reliability-adjusted rates and hospital ranks and assessed changes in rates and ranks between the current and modified approaches. For each comparison, we examined the plots of the rates and ranks to decide whether the modifications had an effect on hospital rankings and determine the relative magnitude of the effects. We also calculated the correlations for the rates and ranks for each comparison.

The mean reliability-adjusted rates and hospital ranks are affected by the choice of the prior distribution. The magnitude of the effect depends on the alternate distribution being tested and the QI. In addition, small hospitals are more likely to be influenced by the prior distribution, which is not surprising given the extent to which small hospitals are pulled toward the mean on average during the reliability-adjustment process. Changing the prior from the normal distribution to a skewed distribution such as a beta or a gamma had a substantial effect on the rates on the right tail of the distribution (high rates or low quality), indicating that a skewed prior distribution may increase the accuracy of estimated rates for outliers. However, the choice between skewed priors (for example, beta versus gamma) has little impact on rates. We also examined the effect of adding hospital characteristics to the prior distribution without changing the normality of the distribution. In most cases, this addition led to substantial separation or grouping of hospital reliability-adjusted rates based on the characteristic added to the distribution. Lastly, the implementation of a formal empirical Bayes framework has little effect on hospital rates and rankings, although the approach has the substantial benefit of clearly and explicitly stating the underlying assumptions, which will allow users to be fully cognizant of all assumptions that underlie their estimates and assist them in making correct inferences that consider these assumptions.

Table ES.6. Recommendations for implementing an empirical Bayes framework

Category	Recommendations
Modification to methods	Regardless of whether AHRQ adopts a formal empirical Bayes framework, clarity of exposition with explicit assumptions stated is crucial to ensuring that users are able to correctly estimate QI rates and draw inferences from the findings. There are substantial changes in the rates from the framework and alternate priors, but there is not enough evidence currently to state if the changes represent improvements. Although the observed distributions of rates for some QIs suggest that the current assumption of a normal distribution should strongly be reconsidered.
Target audiences	Any user that estimates reliability-adjusted rates will benefit from statistical inferences that are based on clearly and explicitly stated assumptions about the prior distribution of the AHRQ QI rates. Under the empirical Bayes framework, AHRQ would be required to modify the software to incorporate the explicit distributional assumptions. In addition, if an alternate prior distribution is ultimately chosen for which there is no explicit form of the posterior distribution, AHRQ will need to build Markov chain Monte Carlo routines into the software so that users can produce results based on simulating posterior distributions of hospital rates. The modification to the software would not reduce the ability of individual hospitals to estimate their hospital rates.
Future analysis	Due to the importance of the prior distributions, we recommend further analysis devoted to better understanding the distributions of hospital rates and assessing whether more appropriate alternate priors exist to the normal distribution. In addition, we recommend further study of approaches to avoid the substantial shrinkage of rates, particularly for small hospitals (such as a multivariate limited translation hierarchical Bayes estimator (Ghosh 2011)).

Key considerations when implementing modifications

The findings provided by the analyses discussed above should be jointly considered with a variety of other factors when deciding whether and how to modify the current approaches to risk and reliability adjustment. In this section, we summarize key considerations for AHRQ when making these decisions, including the conceptual rationale for why a modification is needed, the objectives and needs of various users of the QIs, and additional analyses that will provide evidence to support the decisions.

Because empirical tests cannot definitively establish the relationship between differences in hospital quality and differences in QI rates by hospital type, we recommend that decisions about the role of hospital characteristics in comparisons of QIs be informed by a well developed conceptual rationale. For example, the existence of a clinical basis for differences in the appropriate treatment of conditions presenting at different hospital types might be the basis for concluding that QIs for that condition should not be compared between hospitals of different types. In addition, it is important to consider the suitability (drawbacks and advantages) of different approaches and modifications for users of the QIs with a range of objectives, including hospital quality improvement, direct hospital comparisons across hospital types, and patient decision-making regarding site of care. Although the QIs are not designed and maintained for any one specific user or even specific use, AHRQ can provide added flexibility to the methodological approach for users to better align the approach and the end uses. These more flexible methods would allow users to account for hospital characteristics by modifying or reestimating parameters for instances in which AHRQ deems it appropriate. AHRQ can provide the guidance and tools to facilitate these variations on the current approach (that is, the specifications needed to define eligible cases and adverse events, model specifications to allow reestimation, and clear methods documentation).

Considering the uncertainty surrounding the factors behind differences in QI rates by hospital type, difficult policy decisions will need to be made by the users of the QIs and other quality measures. Ultimately, the decision regarding whether and how to consider hospital characteristics is a policy decision that should be made after carefully considering the objective of the program's use of the QIs, conceptual rationales for the relationships between hospital characteristics and the QIs, and all available empirical evidence regarding the relationships and the appropriateness of various approaches.

Additional research suggested by project findings

The analyses conducted under this project identified further analysis that might improve the accuracy of QI rates used in hospital comparisons. In addition to extensions of the analyses on the AHRQ risk- and reliability-adjustment methods discussed above, possible subjects of this research include: (1) examining factors contributing to differences in QI rates by hospital type, (2) exploring methodological improvements for the composite indicators, (3) exploring a unified risk- and reliability-adjustment approach, and (4) evaluating methods for making inferences about differences in quality.

Examining factors contributing to differences in QI rates by hospital type. It is difficult to uncover evidence regarding the extent to which different factors, such as patient characteristics, processes of care, and structural quality contribute to the differences in QI rates by hospital type. The value of the findings in the analyses discussed in this report and future analyses in providing clear confident recommendations will increase substantially in improving the AHRQ QIs as evidence regarding these factors is uncovered. Modifications can then be designed based on a specific and well understood threat to validity of the rates rather than hypothesized threats. Any analysis that contributes to this understanding should be made a priority for a research agenda regarding risk-adjustment methods. Potential analyses include:

- Sensitivity analyses that inform the likelihood that an unmeasured risk factor could explain the observed differences in rates by hospital type
- Enhanced matched case-control and instrumental variable approaches that attempt to isolate differences in quality from other factors.
- Comparisons of the relationships of QI rates and process measures with hospital characteristics to provide evidence regarding the likelihood that the former is driven by differences in quality rather than risk.
- Examination of the effect of the mass of hospitals with zero rates of adverse events on comparisons of QI rates by hospital type
- Examination of the risk posed by complex patients, such as the risk for burn victims or patients suffering from multiple trauma, and differences in the treatment of these cases by hospital type

A logical extension of the analysis incorporating hospital type indicators in the risk-adjustment models (discussed above) is to examine the effects of incorporating *patient* characteristics contained in discharge data in the QI risk-adjustment models. A critique of the general approach to risk adjustment for quality indicators asserts that patient risk varies by the socioeconomic status (SES) of patients (even after accounting for the demographic and health characteristics of patients in the current QI models), and the proportion of patients with low SES varies by hospital and hospital type. Hence SES would be included as a risk factor in the models to account for unmeasured risk and improve the accuracy of the estimated QI rates. However, as with hospital characteristics, the challenge is determining whether differences in the rate of adverse events for discharges with a characteristic are due to the risk generated by the characteristic or differences in care delivered to patients with the characteristic. In the SES example discussed above, it could be that low-SES patients have an elevated risk regardless of care, or it could be that low-SES patients receive lower quality of care on average (or a mix of both). A simulation analysis following the approach described above could provide evidence regarding how models incorporating SES perform under different explanations of the differences in rates.

Exploring methodological improvements for the composite indicators. Many of the current uses of the AHRQ QIs for hospital comparisons make use of the PSI 90 composite indicator (safety for selected indicators). Because of the importance of the composite indicators in hospital comparisons and the demonstrated relationships with hospital characteristics, a logical extension of AHRQ's QI methods research is to examine the effect of modifications to the current methods on composite indicators. We recommend additional analysis on approaches to improving the accuracy of composite indicators, which include:

- Assessment of the "downstream" effects of modifications to the component indicators discussed above on the composites they comprise

- Examination of modifications to the methods used to construct the composites; in particular, approaches to weighting (for example, weights based on the salience of QIs for a given programmatic objective or positive predictive value of the QIs, which may vary according to hospital type)

Exploring a unified risk- and reliability adjustment approach. Relative to AHRQ, CMS adopts a similar but modified approach to obtain reliability-adjusted estimates of hospital quality. Instead of using a two-stage model in which the risk- and reliability-adjustment steps are separate, a unified hierarchical logistic model is used to fit the discharge-level data. This unified approach shrinks model parameters as part of a single estimation procedure, while AHRQ estimates risk-adjustment parameters without reliability adjustment and then reliability adjusts the resulting risk-adjusted rates in a second stage. There are two important potential advantages of a unified approach. The first is that the statistical uncertainty inherent in risk adjustment propagates naturally through to the final inference of interest in a two-stage approach. The second is that all model parameters are jointly estimated and their covariances are therefore appropriately accounted for in a unified approach. These benefits could lead to more accurate estimation of hospital rates and the uncertainty of those rates for use in making inferences. A practical potential disadvantage of a unified model is that parameter estimation depends on the full data set of raw rates across all hospitals, meaning that each time new data are to be considered, the full national-level model must be rerun. AHRQ's approach, by contrast, fixes these parameters a priori, making it possible for each hospital to estimate their own reliability-adjusted rates. In order to determine whether the statistical advantages of a unified approach suffice to outweigh its practical disadvantages, we recommend specifying and fitting a unified model to AHRQ's discharge-level dataset and comparing the resulting estimates to those from the current approach.

Evaluating methods for making inferences about differences in quality. We also recommend analysis of the ways in which QIs are applied that might be affected by hospital type. For example. we recommend further analysis of methods that fully incorporate statistical uncertainty in inferences based on hospital rates. In a Bayesian framework, these estimates of uncertainty could be used to enhance inferences when making hospital comparisons, such as the "exceedance probability" technique proposed by Ash et al. (2011). In assessing the performance of PSI rates in comparative reporting on patient safety, AHRQ could simulate hospital ranks based on exceedance probabilities and those based on the current or alternate modified approaches. In addition, we recommend that AHRQ study the possible role of stratification or peer grouping of hospitals by their characteristics as an approach to making inferences regarding comparative hospital performance.

This page has been left blank for double-sided copying.

I. INTRODUCTION

The Agency for Healthcare Research and Quality (AHRQ) Quality Indicators™ (QIs) were developed to help states assess inpatient quality of care at hospitals. AHRQ developed three categories (modules) of QIs that estimate rates of different types of adverse events at the hospital level: Patient Safety Indicators (PSIs), Inpatient Quality Indicators (IQIs), and Pediatric Safety Indicators (PDIs). The QIs were soon also used by hospitals to monitor their performance regarding patient safety and mortality. Furthermore, a demand for comparisons of quality between hospitals for various public and private programs led to the risk and reliability adjustment of the QIs. The leveling feature of the risk-adjustment process and the adjustment for the reliability of estimated hospital rates of adverse events facilitated use of the QIs to compare hospital quality in quality-improvement initiatives. For example, the QIs are being used in federal programs to make publicly reported comparisons of hospital quality indicators and payment adjustments based on hospital rankings. The validity of these comparisons depends on the effectiveness of the risk and reliability adjustment of the QIs.

Considering the high profile and high-stakes uses of the indicators in comparative reporting programs, AHRQ has made it a priority to identify threats to the validity of the QIs for use in hospital comparisons. In particular, AHRQ aims to better understand whether the differences across hospital types in QI results that are used to make such comparisons reflect factors other than quality of care. In fact, in the scientific literature and popular press, recent critiques of the use of hospital quality indicators in comparative reporting have focused on whether the current methods support comparisons of quality of care among different hospital types, which deliver care in different settings with different patient populations.

The objective of this project is to make recommendations regarding potential modifications to the AHRQ QI methods for the purpose of improving the accuracy of hospital comparisons. To achieve this objective, we studied the differences in the AHRQ QI rates across hospital types and examined potential modifications to the methods used to estimate the AHRQ QI rates. In particular, we focused on potential modifications to the risk-adjustment and reliability-adjustment methods.[4] Through our review of risk- and reliability-adjustment methods, we identified four specific areas in which opportunities for improvement could be found and identified modifications in each of the areas that could lead to greater accuracy in hospital comparisons using the QI rates. The four areas are: the method used to standardize hospital rates, incorporation of hospital characteristics in risk adjustment, shrinkage (also referred to as smoothing) to targets that vary according to hospital type or the characteristics of the study sample for the reliability-adjusted rates, and implementing a formal empirical Bayes or Bayesian statistical framework for estimating reliability adjusted rates (Bohl et al. 2014; Chen et al. 2014; Jones et al. 2014a, b; Wang et al. 2014). We tested the implementation of the modifications and used the findings in these analyses to support recommendations for potential modifications and identify areas requiring further analysis before more specific recommendations can be made.

[4] Assessing modifications to the discharge- or patient-level variables included in the risk-adjustment models is outside of the scope of this project. We discuss potential extensions of the analyses conducted under this project and logical next steps in each of the four chapters describing the analyses and in the Discussion chapter.

In this report, we summarize each of the four modifications examined as potential improvements to the current risk- and reliability-adjustment methods. We summarize the analyses and make recommendations based on the evidence uncovered in the analyses. We conclude with a discussion of the overarching recommendations supported by the analyses conducted under this project and recommend additional research beyond the scope of this project, which include analyses of the methods used to estimate composite indicators and the overall approach (jointly considering risk and reliability adjustment). In addition, because the approaches used to apply and interpret the QIs for a range of programmatic purposes can have a large impact on the accuracy of hospital comparisons, we recommend next steps for assessing how the QIs are used and interpreted. Furthermore, although there is no specific end use or end user in mind when making the recommendations, we consider the suitability (drawbacks and advantages) of different methodological approaches and modifications for broad classifications of end uses and end users (a hospital aiming to improve quality of care, programs comparing large numbers of different hospital types, and patients making decisions regarding site of care).

A. Methodological approaches targeted for improvement

The primary challenge for improving the risk-adjustment methods examined in the project is the potential that the current methodological approach produces measures of adverse events for hospitals that reflect factors other than the hospitals' performance. Furthermore, these factors could vary by hospital type, potentially misrepresenting the performance of entire classes of hospitals based on estimated QI rates. The primary challenge for improving the reliability-adjustment methods is that hospital rates might be adjusted using information that doesn't accurately reflect the quality of hospitals. It could be that the adjustment doesn't reflect differences in quality by hospital type or that it doesn't reflect quality for the specific group of hospitals being examined by the user. We summarize areas for improvement to the current risk- and reliability-adjustment approaches below and introduce the modifications to these approaches to be tested as improvements in subsequent sections of this chapter.

Approach to risk adjustment

The raw rates of adverse events estimated by the QIs are adjusted based on patient risk factors indicated in patient discharge records to account for factors that increase or decrease a patient's risk for a given adverse event but which are not influenced by the quality of care delivered to the patient (for example, a patient's gender, age, or comorbidities that are present at the time of admission). These factors are used to predict the number of adverse events at each hospital (expected rate) given the hospital's patient risk profile and the estimated relationships between the risk factors and adverse events in the reference population AHRQ uses to estimate the risk-adjustment models. That relationship is estimated by logistic regression over the Healthcare Cost and Utilization Project (HCUP) reference population of all payer discharges from 44 states. Coefficients from these regression models can be used to predict the likelihood for any given patient or set of patients if their characteristics are known, and the predicted likelihoods can be aggregated for hospitals into expected rates of adverse events. A hospital's risk-adjusted rate is then calculated by indirect standardization; that is, the observed rate of adverse events is divided by the expected rate of events, which is multiplied by the reference population rate to arrive at a rate for each hospital. Thus, a hospital's risk-adjusted rate can be interpreted as the performance of a hospital treating its patients relative to a hypothetical average hospital treating patients with the same characteristics.[5] In effect, if one could fully account for all risk factors, the approach levels the playing field across hospitals by adjusting hospitals' rates according to the risk profiles of their patient populations.

A potential concern with the risk-adjustment approach is that indirect standardization might lead to inaccurate comparisons across hospital types. In effect, indirect standardization compares a hospital's actual performance to the expected performance of a hypothetical average hospital. However, the relationship between adverse events and risk factors could vary for some patients by hospital type. By estimating the relationships for an average hospital and comparing a hospital's performance to an average hospital across all hospital types, this approach could miss differences in quality by hospital type within the larger overall set of hospitals.

A second (albeit closely related) potential concern with the current risk-adjustment approach is that differences in hospital rates (including by hospital type) reflect differences in patient risk profiles that are not included in the risk-adjustment models rather than differences in the quality of care delivered to patients. This would be the case if a certain type of hospital is more likely to treat patients with a risk factor that is not included in the QI risk-adjustment models. Hospitals with higher proportions of patients with the risk factor would have more adverse events and higher observed rates, all else being equal. However, their risk-adjusted rates would not be adjusted for having higher proportions of these patients, and their risk-adjusted rates would also be higher, all else being equal. The differences in rates between hospital types would be in part due to the differences in the unmeasured risk factor rather than solely differences in quality of care delivered to patients as intended by the methodological approach. Thus, comparisons of hospitals using these risk-adjusted rates could lead to spurious conclusions regarding the relative quality of hospitals across hospital types.

Approach to reliability adjustment

The AHRQ QI risk-adjusted rates are reliability adjusted to account for uncertainty about an individual hospital's rate arising from the limited information about that hospital's own performance in its discharge records. Through AHRQ's approach to reliability adjustment, the risk-adjusted rates are the weighted average of the hospital's own rate and a reference population rate believed to provide an estimate of the hospital's likely performance in the absence of any information from its own discharges. That estimate is the mean rate from the HCUP reference population.

The weight assigned to the hospital's own rate is the reliability weight, a measure of the reliability of the hospital's estimated risk-adjusted rate. The reliability-adjusted rate is described as shrunken to the reference population rate according to the reliability weight. The reliability weight is estimated following an empirical Bayes framework with a prior distribution estimated from the reference population. The approach assumes that hospital rate variation in that population is normally distributed and its mean is given by the reference population mean. The signal variance and mean of this distribution are the parameters that determine the extent to which the risk-adjusted rates are shrunken (signal variance) and determine the rate to which it is

[5] Another way of interpreting the risk-adjusted rate is as an estimate of adverse events the hospital would experience if it had the average patient population, given the hospital's performance with its actual patient population.

shrunken (mean rate).[6] The accuracy of hospital reliability-adjusted rates depends on the appropriateness of the data used to estimate these; however, for certain applications of the QIs, data other than HCUP reference population may provide more accurate reliability-adjusted rate estimates. In addition, appropriate use of the AHRQ QIs depends on a clear and explicit statement of a statistical framework and the model assumptions within that framework.

The choice of parameters can have a substantial effect on the accuracy of hospital quality comparisons, especially for hospitals that see a limited number of patients and whose rates are shrunken toward the mean rate from the chosen prior distribution to a greater extent on average (Jones et al. 2014a, Bohl et al. 2014). It is important to consider the information available when identifying the most appropriate prior for use in the reliability adjustment process, not only in to avoid the use of a prior that does not match the population to be analyzed but to ensure the use of the most appropriate prior available. We highlight two broad categories of applications for which using information provided in different shrinkage parameters could improve the precision of reliability adjustment leading to more accurate hospital comparisons: (1) the analytic sample of discharges and hospitals of interest is substantially different from the all-payer HCUP reference population used in the current approach (for example, a sample of only Medicare discharges or discharges from a different time period than the reference population) and (2) there are differences in quality of care delivered across hospital types, which is information that could be incorporated in the calculation of reliability-adjusted rates.

Two additional potential improvements in the approaches to reliability adjustment are to : (1) improve the clarity of the exposition of the statistical framework and model assumptions used to estimate reliability-adjusted rates and (2) allow flexibility of the assumptions regarding the distributions of hospital rates to improve the accuracy of estimated hospital rates and their use in comparing hospitals. First, the AHRQ documentation can be improved by providing a detailed description of the modeling approach and explicit identification of the distributional assumptions made in the approach. By adding clarity, AHRQ can help avoid the use of incorrect distributional assumptions, which could lead to errors in inferences based on hospital reliability-adjusted rates. In addition, the assumption in the current approach that hospital rates follow a normal distribution is potentially problematic for two reasons. The validity of the reliability-adjusted rates depends on the validity of the assumption that hospital rates follow a normal distribution; thus, deviations from this assumption will introduce error in the estimation of hospital rates. Second, the assumption of a normal distribution could be unnecessarily restrictive and lead to a reduction in reliability weights and a greater than necessary degree of shrinkage.

B. Opportunities to improve methodological approaches

We identified modifications that could improve the accuracy of hospital comparisons in light of the potential concerns with the current methods described above. We analyzed one

[6] The extent to which a hospital's risk-adjusted rate is shrunken to the prior mean is determined by a reliability weight calculated for each hospital as the ratio of the signal variance (between hospital variance) to the total variance (sum of the signal variance and the noise or within hospital variance). Thus, as the hospital's noise variance increases, the reliability weight decreases; that is, the weight placed on the hospital's own rate decreases and the weight placed on the reference population rate increases. QIs with higher signal variance (or the variation in performance across hospitals) tend to have higher reliability weights, which places a higher weight on hospital's own rates.

potential improvement for each issue identified: two for risk adjustment (approach to the standardization of rates and differences in unmeasured risk) and two for reliability adjustment (shrinking to alternate shrinkage targets and a formal empirical Bayes or Bayesian framework). We introduce the potential improvements below and discuss them in more detail in Chapters IV through VII. Table I.1 summarizes four methodological concerns and the potential improvements to address the concerns examined in this report.

Risk adjustment

Approach to standardization. To assess whether indirect standardization is adequately controlling for case mix, we compared directly standardized rates between hospital types for patients with the same characteristics types to the rates for those two types generated through indirect standardization. This direct standardization approach was possible because patients with almost all sets of characteristics were treated in different hospital types, though in different proportions. We implement this approach to compare performance between hospital types, measured by QI rates, on their care for the same case mix of patients. If differences between types in their directly and indirectly standardized results are substantially different, it indicates that risk adjustment by current methods is not adequate for comparisons of QIs for different hospital types. Direct standardization approaches also have promise for users of the QIs that aim to compare performance to a specific population (for example, high-risk patients, Medicare patients, teaching hospitals, or hospitals within a network) rather than an average patient population from the HCUP reference population.

Incorporating hospital characteristics in risk adjustment models. We assessed an approach to incorporating hospital characteristics in the risk-adjustment models to account for potential unmeasured risk by hospital type. In particular, we assess this modification as a potential improvement to the accuracy of hospital comparisons across hospital types. The primary challenge in assessing the modification is to identify whether changes in hospital rates resulting from the modification reflect an adjustment for differences due to unmeasured risk versus changes that reflect obscuring differences in hospital quality. Without more concrete evidence regarding the degree to which differences in hospital risk-adjusted rates reflect risk versus quality, it is difficult to identify whether the modification reflects an improvement. Because of this uncertainty, we conducted a simulation analysis, in which we examine how assumptions about the role of quality and unmeasured risk in outcome differences by hospital type affect the preferred strategy for addressing them in risk adjustment. For example, if the differences in rates across hospitals were driven by differences in quality in reality, we can assess the potential reduction in the accuracy of hospital rankings from including hospital characteristics. Although more analysis is needed to understand the factors contributing to the differences, we can better understand the ramifications of the modifications given different mixes of two factors (risk and quality) and assess whether the potential benefits of the modification outweigh the potential costs.

Reliability adjustment

Shrinking to alternate targets. We examined the use of different prior distributions and shrinkage parameters for different end uses/users of the QIs as a potential improvement to the reliability adjustment of hospital rates. The hypothesis motivating the modification is that using more appropriate shrinkage parameters and parameters that are more appropriate for the population to be analyzed will lead to more accurate estimates of hospital performance (reliability-adjusted rates) and more accurate comparisons of hospitals based on these estimates. We examined two broad cases in which reestimating the shrinkage parameters could lead to such improvements. First, to examine potential improvements for users who are estimating hospital rates for a population that differs from the reference population (for example, an external sample of Medicare fee-for-service discharges from Medicare claims data), we investigated two modifications: (1) reestimating the shrinkage parameters using the analytic sample and (2) reestimating the shrinkage parameters and reestimating the risk-adjustment models. Second, to examine potential improvements to address differences in quality by hospital type, we investigated reestimating separate shrinkage parameters for the different hospital types; in effect, this shrinks hospital rates to the mean of their hospital type rather than the overall mean.

Incorporating a formal empirical Bayes or Bayesian framework. We identified several parts of the technical specification of the current AHRQ QI reliability-adjustment methodology that would benefit from further statistical exposition. First, AHRQ could achieve increased clarity of the methodology under a formal empirical Bayes or Bayesian statistical framework. In the Bayesian framework, for example, all statistical inferences are made through the QI rates' posterior distributions, which are explicitly derived from a prior distribution and a likelihood for the risk-adjustment rates. Through these two distributions, we formally articulate each assumption that underlies the statistical model. A key benefit is that explicit statement of assumptions enables users to better assess the model's suitability for their application. Second, the issues identified earlier regarding the assumption that hospital rates follow a normal distribution (potential misalignment with true distributions of hospital rates and the restrictiveness of the assumption) motivates us to consider alternative prior distributions that may improve the estimation of the AHRQ QI reliability-adjusted rates. In so doing, we investigate prior distributions that may provide a better fit to the data.

Table I.1. Summary of modifications tested

Area of analysis	Methodological challenge	Potential improvement
Risk adjustment		
Standardization approach	The current approach of estimating indirectly standardized rates might not adequately capture the effect of differences in case mix by hospital type.	We compare to directly standardized rates to assess the current approach and discuss possible instances in which direct standardization could be an improvement.
Risk-adjustment model specifications	There could be differences in unmeasured risk by hospital type.	We examine the inclusion of indicators for hospital type in the risk-adjustment models.
Reliability adjustment		
Shrinking to alternate targets	The current shrinkage target (mean of the 44-state HCUP reference population) might not be appropriate for some populations. In addition peer group mean rates could provide more appropriate shrinkage targets.	We examine alternate shrinkage targets for various populations as well as shrinking to peer group means.
Empirical Bayes framework	The current approach lacks a formal statistical framework with explicit assumptions tied to the framework. In addition, the assumptions of normal distributions in the methods could be overly restrictive and possibly misaligned with the data on hospital quality.	We apply empirical Bayes and Bayesian frameworks and examine the effect of alternate distributional assumptions.

Next, we briefly describe the data sources, selection of QIs and hospital characteristics to be examined in the various analyses, and construction of the analytic file to be used in the various analyses. In Chapter III, we summarize the findings of two initial analyses, a review of the literature focusing on the relationships between the QIs and hospital characteristics and an exploratory data analysis (EDA), in which we systematically examine the relationships between all risk-adjusted PSIs, IQIs, and PDIs and a wide range of hospital characteristics. In Chapters IV through VII, we summarize the four analyses of modifications to the risk- and reliability-adjustment methods that aim to improve the accuracy of comparisons across hospital types using QI rates. We conclude with a discussion of the recommendations that follow the findings of the literature review, EDA, and the four methods analyses and suggest areas for future study to continue improving all aspects of the AHRQ QI methodological approach.

This page has been left blank for double-sided copying.

II. DATA AND THE AHRQ QI SOFTWARE

The objective in identifying the data sources was to obtain discharge data that are as close as possible to the discharge data used to create the current AHRQ software so that we can test modifications to the methods and isolate the effect on the results apart from any effects that are the results of differences between the data sources. The analytic file is composed of discharges originating from the State Inpatient Databases (SID), HCUP, coordinated by AHRQ, which is the same source, in the same format, and covers the same period as the file used by AHRQ to estimate the risk-adjustment models. The discharge records that are the basis for the analytic file are obtained from 12 states in 2009 and 2010, a subset of the SID, and thus represent a subset of the data used in development of the current AHRQ QI risk-adjustment models (reference population).[7] The SID contain inpatient discharge abstracts for patients of all ages and for all payers admitted to all community hospitals in participating states.[8]

We added information to these discharges regarding the characteristics of hospitals delivering care to the patients so that we can study the relationships between QI results and hospital characteristics and test if the modifications could improve comparisons across hospital types. We merged hospital-level information from several sources onto each discharge using hospital IDs. Many of the hospital characteristics were obtained from the 2010 American Hospital Association Survey Database; other sources include the 2010 Centers for Medicare & Medicaid Services (CMS) Impact File, CMS Certification Numbers, 2013 United States Department of Agriculture Economic Research Service Rural-Urban Continuum Codes, and aggregate information from the discharge data, such as the percentage of discharges for various primary payers (Medicaid, Medicare, and uncompensated care).

In the first use of the analytic file, the EDA, we examined the relationships between adverse events and hospital characteristics at the discharge level. We also used the discharge data to create hospital-level results according to the current QI methods using the specifications in the AHRQ QI software without modification. The resulting hospital-level data were the basis of the hospital-level comparisons by hospital type examined in the EDA. In addition, these discharge- and hospital-level data serve as a baseline for the comparisons of the results from the four modified methods. Generally speaking, we modified the methods and tested the effects of the modifications by comparing the results by hospital type to those generated using the unmodified current methods. For more information on the data sources and construction of the analytic file, see the relevant sections of the EDA and the four methods analyses (Bohl et al. 2014; Chen et al. 2014; Jones et al. 2014a, b; Wang et al. 2014).

In the EDA, we examined all risk-adjusted PSIs, IQIs, and PDIs and a large set of hospital characteristics. The hospital characteristics can be grouped into three categories based on their

[7] We would like to acknowledge the HCUP Data Partners: Arizona Department of Health Services, Arkansas Department of Health, California Office of Statewide Health Planning and Development, Florida Agency for Health Care Administration, Iowa Hospital Association, Kentucky Cabinet for Health and Family Services, Maryland Health Services Cost Review Commission, Massachusetts Center for Health Information and Analysis, Nebraska Hospital Association, New Jersey Department of Health, New York State Department of Health, and Washington State Department of Health.

[8] For more information on the SID and HCUP, see http://www.hcup-us.ahrq.gov/sidoverview.jsp.

hypothesized relationships with estimates of quality of care. Structural characteristics include organizational and operational characteristics of hospitals related to levels of resources and experience (for example, teaching status and bed size). Aggregate patient hospital characteristics represent factors that might indicate unmeasured individual patient risk or factors that directly affect hospital resources (for example, disproportionate share status [DSH]). Market and local area characteristics represent factors external to the hospital that can affect the primary population served by hospitals, and thus unmeasured individual patient risk (for example, critical access hospitals [CAHs]).

However, given the large number of QIs and hospital characteristics examined in the larger project, it is not feasible to test the modified models for every combination of QI and hospital characteristic and synthesize the results for presentation in a report format. Thus, we selected a subset of QI and hospital characteristic combinations for closer examination in subsequent analyses. We selected this subset with the objective of maximizing the generalizability of the findings in this report to the full set of QIs. Hence, we chose QIs that represent a range of clinical properties (such as PSIs addressing continuity of care as well as technical care) and statistical properties (such as including QIs that measure rates of relatively rare events as well as more common events). In addition, because composite indicators are used to make hospital comparisons in a range of federal programs (such as the CMS's Inpatient Quality Reporting program and Hospital Value-Based Purchasing program), we prioritized QIs that have the largest weights in the calculations of hospital composite values. Although not all QIs showed the same relationships with the hospital characteristics examined in the EDA, patterns emerged across the QIs within the three modules which suggests that including a subset of QIs from each module is likely to produce results that represent the rest of the QIs. Although the specific combinations of QIs and hospital characteristics varied slightly by analysis, the primary relationships examined are between the QIs and hospital characteristics listed in Table II.1.

Table II.1. AHRQ QIs and hospital characteristics included in the analysis

Quality indicators	Hospital characteristics
• PSI 06 Iatrogenic Pneumothorax Rate • PSI 12 Postoperative PE/DVT Rate • PSI 13 Postoperative Sepsis Rate • PSI 14 Postoperative Wound Dehiscence Rate • PSI15 Accidental Puncture or Laceration Rate • IQI 15 Acute Myocardial Infarction Mortality Rate • IQI 16 Heart Failure Mortality Rate • IQI 19 Hip Fracture Mortality Rate • IQI 20 Pneumonia Mortality Rate • PDI 01 Accidental Puncture or Laceration Rate • PDI 10 Postoperative Sepsis Rate • PDI 12 Central Venous Catheter-Related BSI Rate	• Number of licensed hospital beds: broken into indicators for bed size quartiles • Teaching hospital status: indicator for major/minor teaching status • Disproportionate share (DSH) status: indicator for greater than 15 percent of patient populations composed of disproportionate share patients • Critical access hospital (CAH) status: indicator of designation as a critical access hospital

Notes: PE/DVT = pulmonary embolism or deep vein thrombosis; BSI = blood stream infection.

We selected the hospital characteristics that demonstrated an empirical relationship with QI results in the literature, have a strong conceptual rationale for such a relationship, and are important in the policy context of making hospital comparisons: number of licensed hospital beds (broken into indicators for bed size quartiles), teaching hospital status (indicator for major/minor teaching status), DSH status (indicator for greater than 15 percent of patient populations composed of disproportionate share patients), and CAH status (indicator of designation as a critical access hospital). In addition, based on the EDA, we selected hospital characteristics associated with statistically significant differences in QI rates that are fairly consistent across QIs within the three QI modules. The selected hospital characteristics are also representative of other relevant characteristics omitted from these analyses (for example, hospitals that meet the criteria for disproportionate share status typically also have relatively high proportions of uncompensated care), thus maximizing the generalizability of the findings (Jones et al. 2014a; Dy et al. 2013). Furthermore, the selected characteristics include structural characteristics of hospitals (bed size and teaching status) as well as characteristics of aggregate patient populations (DSH) and the local settings in which the hospitals are located (CAH). For more information on the selection and definition of the QIs and hospital characteristics, see the relevant sections of the four methods analyses (Bohl et al. 2014; Chen et al. 2014; Jones et al. 2014a, b; Wang et al. 2014).

This page has been left blank for double-sided copying.

III. OVERVIEW OF DIFFERENCES IN QI RESULTS BY HOSPITAL TYPE

Before we discuss modifications to the QIs, we summarize results from two analyses that increased our understanding of the differences in QI rates across hospital types, and helped us to identify the modifications, hospital types, and QIs we selected for in-depth study. First, we reviewed the literature investigating differences in hospital QI rates by hospital characteristics and the methods used to estimate QI rates. We followed this review with an extensive EDA that systematically examines the relationships between hospital QI rates and hospital characteristics using a 12-state sample of hospital discharge data.

A. Literature review

The primary objectives of the literature review were to better understand the relationships between AHRQ QI results and hospital-level characteristics that could influence the estimation of the results and to identify potential threats to the validity of the results that could be addressed either in the AHRQ QI methodology or in recommendations about reporting. To meet these objectives, we examined literature that addresses the associations between hospital characteristics and results for all risk- and reliability-adjusted AHRQ PSIs, IQIs, and PDIs. We also reviewed the literature to identify modifications to the statistical methods used to estimate the QIs (for example, reliability adjustment, risk adjustment, and construction of composites) that could impact the estimates and lead to more accurate comparisons of hospital quality.

We conducted a search of both the published literature and the gray literature from January 2000 to December 2012. The final review included 44 studies reported in 45 articles focusing on a wide range of combinations of the QIs and hospital characteristics. None of the studies included a systematic review of all QIs or all QIs within a module. Each study focused on one QI or a small subset of the QIs and one or a small number of hospital characteristics.

A relationship between a QI and a hospital characteristic was assessed based on the consistency in the direction of the associations (positive or negative) and the statistical significance of the associations across studies. There were no definitive relationships between the QIs and hospital characteristics (that is, no relationships were completely consistent in direction and statistical significance of the association) across studies examined in the literature review. Many of the studies found that the estimated associations between the QI rates and hospital characteristics studied were not statistically significant. However, for some combinations of the QIs and hospital characteristics, the associations were either in the same direction or not statistically significant. For example, for bed size and 8 of 15 PSIs examined, approximately half of the studies found that greater number of beds was associated with higher odds of patient safety events, whereas half of the studies found no statistically significant association; there were no statistically significant associations for bed size and the other 7 PSIs. Surprisingly, there was little mention of what factors might contribute to the observed associations in the studies reviewed.

In addition, the review did not uncover any studies that directly assess the methodological approach AHRQ uses to estimate the QIs results. However, we conducted an informal review of studies focusing on methods relevant to the QIs and studies that helped us to identify possible improvements to the current approach. Thus, a logical extension of the review is to expand the search to include a systematic review of studies examining methods relevant to but not directly focused on the QI methods. For the detail findings of the literature review, see Dy et al. (2013).

B. Exploratory data analyses

The primary objective of the EDA was to systematically and comprehensively examine the differences in QI results by hospital type (for all risk- and reliability-adjusted PSIs, IQIs, and PSIs). First, we estimated the bivariate relationship between each QI and each hospital characteristic, and compared our findings across QIs and hospital characteristics. We examined whether there are consistent patterns in the relationships across the QIs within and across the three modules and for groups defined by related hospital characteristics (such as teaching hospitals and high-volume hospitals); related QIs (for example, all QIs related to surgical procedures); and types of QIs (observed, risk-adjusted, and reliability-adjusted rates). Then, we estimated multivariate relationships between each QI and the array of hospital characteristics to identify the hospital characteristics that are most strongly associated with differences in QI rates (in terms of magnitude and statistical significance of the associations). In addition, we analyzed the variation in several factors hypothesized to explain differences in QIs by hospital type: differential coding practices, patient risk, and the role of volume in the calculation of hospital reliability-adjusted rates. We tested whether these factors vary by the same hospital characteristics that demonstrate relationships with QI rates; an indication that these factors contribute to differences in QI rates. The final objective of the EDA was to identify possible modifications to current QI methods to be tested in subsequent analyses.

The most common pattern observed in comparing the bivariate relationships of QI rates and hospital characteristics is that many hospital characteristics have one association with the patient safety QIs (PSIs and PDIs) but the opposite association with IQIs. Several structural and market/local area characteristics exhibited this pattern (high volume, teaching hospitals, high nurse staffing ratios, and, for PDIs, children's hospitals) as did two characteristics that describe the location or market of hospitals (non-CAHs and hospitals located in urban settings).[9] The same relationships between QIs and hospital types was observed in the multivariate analysis, except that teaching hospitals tended to have higher IQI rates than nonteaching hospitals in the multivariate analysis, but lower rates in the bivariate analysis. Therefore, after volume and other hospital characteristics were considered, teaching hospitals had higher rates on average for all three QI modules studied in the EDA. In addition, hospital bed size was the most consistent (in terms of statistically significant associations with QI rates) and strongest (in terms of the magnitude of the associations) predictor of hospital rates in the multivariate analysis; whereas many of the other hospital characteristics (particularly those highly correlated with volume, such as teaching status and urban location) were not as strong. The findings when examining the composite indicators were largely consistent with the results discussed above for the individual QIs.

The pattern of higher PSI and PDI rates but lower IQI rates could be explained by differences in quality of care, differences in coding of diagnoses, or the statistical properties of

[9] We observed the same relationships between QI rates and volume using two measures of hospital volume, bed size and the number of discharges. We use bed size in the reporting of results for most of the analyses because it is the measure of volume most often reported in the relevant literature and because of the similarity in results in the EDA for bed size and the number of discharges.

the QIs in the three modules. First, hospitals that provide better quality of care may be keeping their patients alive longer (lower IQI rates), increasing the opportunity for patient safety events captured by the PSIs or PDIs. Similarly, differences in length of stay and transfer patterns across hospital types could affect QI rates; however, QIs in all three modules would likely be affected in the same direction (that is, shorter lengths of stay or more frequent transfer patterns would likely reduce PSI, IQI, and PDI rates), and an initial analysis controlling for aggregate hospital length of stay did not attenuate this pattern in the results. A second possible explanation is that there could be differences in the relative reliability of reported primary diagnoses (used to identify cases for inclusion in IQI rates) as opposed to secondary diagnoses (used to identify PSI and PDI numerators) across hospital types. Finally, it is possible that extreme rates—either high or low—at small hospitals are driving these patterns rather than signaling true differences in quality; that is, noise or chance is contributing to these patterns because small hospitals are more likely to have rates equal to zero for PSIs and PDIs and more likely to have rates near 100 percent for IQIs, all of which have very low estimated reliability on average. Further analysis is needed to better understand the factors contributing to the different relationships for PSIs and IQIs. The uncertainty surrounding the factors supports the inclusion of both PSIs and IQIs in any assessment of modifications to the methods, as the modification could have a different effect on indicators in the two modules, leading to different recommended modifications by module.

We also observed, when comparing the bivariate relationships between QIs and hospital characteristics, that some hospital types had higher rates (lower quality) for indicators in all three modules. This pattern is particularly true for aggregate patient characteristics, such as the hospital's DSH status, its proportion of Medicaid discharges, and its proportion of Medicare discharges. The relationships are particularly consistent and strong for DSH, which also demonstrated consistently higher rates in the multivariate regression analysis. The multivariate analysis also exhibited a consistent positive relationship between a hospital's proportion of uncompensated care and higher rates for PSIs and PDIs, whereas the relationship of uncompensated care to QI rates was not consistent in the bivariate analysis. Regarding DSH and uncompensated care, the finding could reflect that these hospitals are often safety net hospitals and have lower resources and possibly less healthy patients on average. The relationship could reflect a correlation between socioeconomic status (SES) and elevated patient risk in combination with higher proportions of patients with lower SES being treated at these hospitals. The negative relationship between SES and health care outcomes has been well documented in the literature (Dy et al. 2013). In addition, it is not surprising that DSH have the same relationships with QI rates as hospitals with high proportions of Medicaid, Medicare, or uncompensated care discharges, given that hospitals qualify for DSH status based on these factors.

The EDA did not uncover evidence to support or refute the hypotheses that differences in coding practices, differences in risk, or the volume–QI relationship account for the differences observed in hospital QI rates by hospital type. First, although predicted rates for PSIs and PDIs are higher for hospital types with higher observed rates, the predicted rates for the IQIs and hospital characteristics are inconsistent. Second, it is feasible that volume is driving much of the variation in QI results, even for other hospital characteristics through their strong correlation with volume (for example, teaching hospitals tend to have higher volumes than nonteaching hospitals). However, the multivariate analysis suggests that after accounting for volume, the QI rates still vary slightly across other hospital types. In fact, several structural and aggregate patient characteristics demonstrate relationships with QI rates in multiple QI modules, particularly for the PSIs and IQIs. Lastly, there is no consistent evidence to suggest that the hospital types with higher PSI and PDI rates and lower IQI rates vary in their coding of diagnoses and POA information.[10,11]

The EDA results helped us determine the hospital characteristics and the QIs to be examined in the next phase of our analysis. We selected for follow-up hospital types exhibiting statistically significant associations in the same direction (that is, positive or negative) with subsets of the QIs (for example, across the majority of PSIs). For example, we examined teaching status and hospital bed size in each of the four methods analyses, as they exhibited consistent relationships with many of the QIs and the intriguing pattern that they have one relationship with PSIs/PDIs and another with IQIs (along with being of interest to policymakers focusing on hospital quality measurement).Also, in order to test associations of varying strength, we selected PSI 12 (postoperative pulmonary embolism/deep vein thrombosis [PE/DVT]) and IQI 20 (acute myocardial infarction [AMI] mortality) for our analysis of incorporating hospital characteristics in risk-adjustment models; PSI 12 exhibited particularly large associations with the hospital characteristics examined, whereas IQI 20 exhibited consistent but small associations. For a detailed exposition of the EDA findings, see Jones at al. (2014a).

[10] The one exception to this finding occurs when examining coding of diagnoses at CAHs. There is a substantial difference in the number of diagnosis codes reported by CAHs and non-CAHs. CAHs tend to report fewer diagnoses and also have lower PSI and PDI rates. Exploratory analyses also detect potential for anomalous present-on-admission coding at CAHs.

[11] Although the evidence in the EDA does not support the hypothesis, if the differences are due to factors such as differential coding or patterns of length of stay and transfers, the potential solutions or modifications to the methods will likely have more to do with coding guidance and the incorporation of preadmission and postdischarge information, rather than modifications to the risk-adjustment and reliability-adjustment methods. These two potential modifications are beyond the scope of this project but are nonetheless considered as potential causes of the differences in QI rates by hospital characteristics.

IV. ASSESSING BENEFITS AND LIMITATIONS OF INDIRECT STANDARDIZATION FOR RISK ADJUSTMENT

A. Methodological challenge

The current risk-adjustment methodology used to estimate the AHRQ QIs is indirect standardization. By indirect standardization, a hospital is compared to a hypothetical average hospital in the reference population based on their relative outcomes for patients similar to those of the hospital. For example, the AHRQ software contains a risk-adjustment model that was estimated over the HCUP reference population, representing all payers. For a given hospital's patients, the estimated model produces an expected rate of adverse events for an average hospital treating those patients. The hospital's observed rate is compared to the expected rate, which answers the question, "How do the hospital's patients fare compared with our expectation of how they would fare at an average hospital in the reference population?" By contrast, when direct standardization is used, the question answered is "How would a predetermined set of patients (for example, those in the reference population) have fared had they been treated at this hospital?

A possible limitation of the current risk-adjustment methodology is that it does not account for the way that patient characteristics can be correlated with hospital characteristics. The case mix of patients can vary across different types of hospitals and the current risk-adjustment method might not accurately estimate disease burden (Kolfschoten et al. 2011; Friese et al. 2010). In particular, a model-based approach may not adequately account for lack of overlap in patient case mix between the hospital types. For example, teaching hospitals typically treat more severely ill patients than nonteaching hospitals (Khuri et al. 2001). If teaching hospitals have better outcomes, the result could understate the relationship between severity of illness and the difference in outcomes we might observe if all patients were treated in teaching facilities, compared to their outcomes if they were treated in nonteaching facilities. Therefore, the variation in QI rates by hospital characteristics could be related to either true differences in health care quality or inadequate adjustment for patients' case mix by indirect standardization. The goal of our investigation was to assess whether indirect standardization adequately controls for patient risk factors between hospitals of different types.

B. Potential improvement

Our study assessed the performance of indirect standardization for risk adjustment, in the context of making comparisons of hospital risk-adjusted rates between different types of hospitals. We studied the effect of standardization methods by comparing the differences in risk-adjusted rates between hospital types when the rates are directly standardized with differences in rates when indirectly standardized. By direct standardization, we compare exactly the same types of patients, based on their risk factors, between hospital types. This approach enables us to compare performance between hospital types, measured by QI rates, on their care for the same case mix of patients. If the relationship between the QI and hospital type differs between directly and indirectly standardized results, it might be because the current risk-adjustment methodology does not adequately remove variation due to patient risk factors in comparisons of QIs between hospital types (Fleiss et al. 2003). In such cases, further analysis is needed to determine if a modification to the risk-adjustment models could improve the comparisons of QI rates by hospital type using indirectly standardized rates.

In our approach, a change in the variation in QI rates between types under direct standardization is an indication that indirect standardization does not adequately adjust for case mix between hospital types. If we find such a change, we may recommend a modification to the indirect standardization approach or that direct standardization be adopted for certain applications of the QIs (for example, when comparing performance within a set of hospitals of a particular type or peer group). Similar recommendations can be found in the literature. For example, Silber et al. (2014) suggest that hospital administrators might be interested in assessing how well their hospitals compare to other hospitals that treat the same patients.

Analytic approach

Direct standardization was achieved by stratifying and matching all discharges according to their sets of risk factors (risk profiles), which removes all variation due to case mix in estimating the risk-adjusted rates. Notably, direct standardization is a model-free statistical method for risk adjustment. Adjusting in this way, we first account for differential case mix by including only patient profiles seen at both types of hospitals; that is, we excluded patients with risk profiles that were not common between hospital types. After stratification, we confirmed that the same set of risk profiles was populated for each hospital type and ensured that between hospital types, the distribution of risk profiles was exactly the same (that is, zero observed differences). The latter was achieved by weighting the risk profiles so that they occur equally across hospital types; in effect, the weighting produces comparisons of rates between hospital types that are for the average discharge populations represented by the hospital types. We also examined excluded discharges with risk profiles that were not represented at one or the other hospital type.

To calculate the directly standardized rates, first we calculated QI rates within each risk profile, that is, the stratum of discharges containing the same case mix of patients (calculated separately for the two hospital types to be compared). These stratum-specific rates were combined, weighted by the number of combined discharges in the risk profile between the two hospital types, to produce the two overall risk-adjusted QI rates, one for each hospital type. These rates correspond to the denominator-weighted average of hospital risk-adjusted rates by hospital type, whereby the denominators are the frequencies of discharges represented by risk profiles in the population.

Differences by hospital type in these directly standardized results were compared with differences in the risk-adjusted rates produced by the AHRQ QI software (version 4.4), which is based on indirect standardization through a logistic regression. We calculated two 95 percent confidence intervals (CIs) for differences in risk-adjusted rates between hospital types: one CI produced by direct standardization; and the other by indirect standardization. If the 95 percent CI for the difference in rates between two hospital types included zero, then the two hospital types were not considered to be statistically different by the given standardization method. We compared whether the differences by hospital type were statistically significant for the rates calculated by indirect and direct standardization. A change in the statistical significance of a difference by hospital type when we estimated rates by direct standardization instead of indirect is evidence that risk-adjustment may be improved, either through direct standardization or changes to model specifications to better support comparisons of QI rates across hospital types using indirect standardization (for instance by incorporating discharge-level factors that vary by hospital type and are correlated with increased patient risk, such as SES).

As described above, we focused the analysis on a subset of the risk-adjusted PSIs, IQIs, and PDIs that represent a range of clinical properties (the types of adverse events addressed by QIs) and statistical properties (for example, QIs that capture events that are relatively rare as well as those that are more common) and QIs that are key components of the PSI, IQI, and PDI composite indicators (that is, they carry the largest weights in the calculation of the composite indicators). We focused on two hospital characteristics that have drawn increased attention in the literature and by policymakers: discharge volume (measured by bed size) and teaching status. For a more detailed account of this analysis, see Chen et al. (2014).

C. Findings

Inpatient Quality Indicators

In general, patient characteristics were similar between teaching and nonteaching hospitals, and between large and small hospitals before stratification. Despite the significant *p*-values due to large samples sizes, the differences were less than 0.1 standard deviations apart, indicating minimal differences in most comparisons. However, certain patient characteristics varied noticeably (standardized difference greater than 0.1) between different types of hospitals and by QI. Compared with nonteaching and small hospitals, patients of teaching and large hospitals in the denominator population for mortality related to AMI, heart failure, and pneumonia were younger. Furthermore, AMI and hip fracture patients of teaching and large hospitals were more likely to have transferred from other hospitals. Even considering these differences, for comparison of directly standardized rates by teaching status and hospital size, more than 99.6 percent of patients were included in the stratification; less than 0.4 percent were excluded because they had risk profiles not represented in both hospital types. In addition, excluded cases represented patients with relatively rare combinations of risk factors.

As demonstrated in Table IV.1, compared with nonteaching hospitals, teaching hospitals had lower indirectly standardized rates for AAA repair mortality (IQI 11) and pneumonia mortality (IQI 20) on average. The differences in indirectly standardized rates by teaching status were not statistically significant for AMI mortality (IQI 15) and heart failure mortality (IQI 16). Similar patterns were found in their directly standardized rates for these IQIs. In contrast, teaching hospitals had higher indirectly standardized rates for hip fracture mortality (IQI 16) on average, but the difference in directly standardized rates was not statistically significant.

Compared with small hospitals, large hospitals had lower indirectly standardized mortality rates for AAA repair, AMI, heart failure, and pneumonia on average. These differences were also demonstrated in directly standardized rates, except for the rates of AAA repair mortality, for which the directly standardized difference was not statistically significant.

Table IV.1. Observed, indirectly standardized, and directly standardized IQI rates, by hospital type (per 100 discharges)

		Teaching status			Bed size		
		Teaching	Nonteaching	Difference [95% CI]	Large[a]	Small[a]	Difference [95% CI]
IQI 11	Observed	4.50	5.06	-0.57	4.51	5.37	-0.86
	Indirect	4.51	5.23	-0.73* [-1.22, -0.24]	4.61	5.42	-0.80* [-1.35, -0.26]
	Direct	4.49	5.05	-0.56* [-1.06, -0.06]	4.61	5.17	-0.56# [-1.21, 0.08]
IQI 15	Observed	5.84	6.95	-1.11	5.95	7.36	-1.41
	Indirect	6.40	6.44	-0.04 [-0.19, 0.12]	6.33	6.57	-0.25* [-0.41, -0.08]
	Direct	6.38	6.45	-0.07 [-0.22, 0.08]	6.32	6.57	-0.26* [-0.45, -0.07]
IQI 16	Observed	3.30	3.44	-0.14	3.29	3.51	-0.23
	Indirect	3.38	3.33	0.05 [-0.03, 0.12]	3.28	3.47	-0.19* [-0.28, -0.10]
	Direct	3.43	3.35	0.08 [-0.01, 0.16]	3.31	3.49	-0.18* [-0.26, -0.1]
IQI 19	Observed	2.86	2.61	0.25	2.71	2.69	0.02
	Indirect	2.83	2.62	0.20* [0.05, 0.36]	2.67	2.72	-0.05 [-0.2, 0.1]
	Direct	2.73	2.65	0.08# [-0.05, 0.20]	2.63	2.77	-0.14 [-0.29, 0.02]
IQI 20	Observed	4.05	4.18	-0.13	4.09	4.18	-0.09
	Indirect	3.97	4.23	-0.25* [-0.35, -0.15]	3.97	4.31	-0.34* [-0.44, -0.25]
	Direct	3.95	4.23	-0.28* [-0.36, -0.19]	3.93	4.34	-0.42* [-0.52, -0.32]

Sources: Mathematica Policy Research analysis of 12 SID from January 2009 to December 2010; fiscal year (FY) 2010 AHA Survey Database.

Note: 95 percent CIs apply only to standardized rates. Teaching hospitals are defined as major or minor teaching hospitals in the 2010 AHA Survey Database.

[a] Large hospitals are those in the quartile with the largest number of beds (more than 271). Small hospitals are those in the quartile with the smallest number of beds (less than 56).

* Denotes that the difference in rates by hospital type is statistically significant at the 5 percent level.

# Denotes that comparing directly standardized rates by hospital type leads to a different conclusion regarding whether the differences are statistically significant.

IQI 11 = abdominal aortic aneurysm repair mortality; IQI 15 = acute myocardial infarction mortality; IQI 16 = heart failure mortality; IQI 19 = hip fracture mortality; IQI 20 = pneumonia mortality.

Patient Safety Indicators

Like the IQIs, patient characteristics for the PSIs were largely similar between teaching and nonteaching hospitals, and between large and small hospitals before stratification. However, teaching and large hospitals treated more patients transferred from other hospitals for postoperative PE/DVT. Teaching hospitals also had younger patients in their discharge populations for iatrogenic pneumothorax, postoperative PE/DVT, and accidental puncture or laceration. Through stratification, we were able to include 99.6 percent or more of patients for all examined PSIs except postoperative PE/DVT (94 percent). After stratification, patient characteristics were identical between different types of hospitals. The excluded cases represented patients with relatively rare combinations of risk factors.

Teaching hospitals had higher indirectly standardized rates of iatrogenic pneumothorax (PSI 06), postoperative PE/DVT (PSI 12), and accidental puncture or laceration (PSI 15) than nonteaching hospitals on average. The difference in indirectly standardized rates by teaching status on average was not statistically significant for postoperative wound dehiscence (PSI 14). These patterns in PSI rates remained using directly standardized rates. For the postoperative PE/DVT rate, because teaching and large hospitals experienced notably smaller risk-adjusted rates after excluding unmatched discharges, the differences between hospital types were smaller in magnitude; however, the differences were still statistically significant.[12]

Similarly, large hospitals had higher indirectly standardized rates of iatrogenic pneumothorax, PE/DVT, and accidental puncture or laceration on average when compared to small hospitals. In addition, the difference in indirectly standardized rates by bed size on average was not statistically significant for postoperative wound dehiscence. Similar patterns were found after applying direct standardization with the exception of rates of accidental puncture or laceration; the difference in directly standardized rates by bed size on average was not statistically significant, although the difference almost met the standard of statistical significance using the 95 percent CI.

[12] It could be interpreted that teaching and large hospitals have substantially different discharge populations for postoperative PE/DVT; however, this difference is based on all interactions of the PSI 12 risk factors (that is, all combinations of risk factors are stratified and matched), which may not be clinically meaningful.

Table IV.2. Observed, indirectly standardized, and directly standardized PSI rates, by hospital type (per 1,000 discharges)

		Teaching status			Bed size		
		Teaching	Nonteaching	Difference [95% CI]	Large[a]	Small[a]	Difference [95% CI]
PSI 06	Observed	0.48	0.35	0.12	0.45	0.33	0.12
	Indirect	0.46	0.37	0.09* [0.07, 0.12]	0.44	0.36	0.08* [0.05, 0.10]
	Direct	0.45	0.36	0.09* [0.07, 0.11]	0.42	0.35	0.07* [0.04, 0.10]
PSI 12	Observed	6.80	4.81	1.99	6.57	4.50	2.07
	Indirect	6.57	4.98	1.59* [1.47. 1.71]	6.42	4.66	1.77* [1.64. 1.91]
	Direct	4.73	3.81	0.91* [0.79, 1.04]	4.59	3.51	1.08* [0.93, 1.23]
PSI 14	Observed	1.87	1.99	-0.11	1.89	1.99	-0.10
	Indirect	1.93	1.92	0.01 [-0.17, 0.19]	1.92	1.93	-0.01 [-0.19, 0.18]
	Direct	1.92	1.91	0.01 [-0.17, 0.19]	1.91	1.91	0.00 [-0.21, 0.21]
PSI 15	Observed	3.13	2.05	1.08	2.71	2.14	0.57
	Indirect	2.64	2.28	0.35* [0.29. 0.42]	2.47	2.41	0.06* [0.00. 0.12]
	Direct	2.67	2.29	0.39* [0.34, 0.44]	2.46	2.43	0.03# [-0.03, 0.09]

Sources: Mathematica Policy Research analysis of 12 SID from January 2009 to December 2010. FY 2010 AHA Survey Database.

Note: 95 percent CIs apply only to standardized rates. Teaching hospitals are defined as major or minor teaching hospitals in the 2010 AHA Survey Database.

[a] Large hospitals are those in the quartile with the largest number of beds (more than 271). Small hospitals are those in the quartile with the smallest number of beds (less than 56).

* Denotes that the difference in rates by hospital type is statistically significant at the 5 percent level.

# Denotes that comparing directly standardized rates by hospital type leads to a different conclusion regarding whether the differences are statistically significant.

PSI 06 = iatrogenic pneumothorax; PSI 12 = postoperative pulmonary embolism or deep vein thrombosis; PSI 14 = postoperative wound dehiscence; PSI 15 = accidental puncture or laceration.

Pediatric Quality Indicators

Patient characteristics for the PDIs differed markedly by hospital characteristics before stratification. Compared with nonteaching and small hospitals, teaching and large hospitals performed more procedures during each individual hospitalization at risk for accidental puncture or laceration, and treated more patients who were transferred from other hospitals and at high-risk for postoperative sepsis and central venous catheter-related bloodstream infection. In direct standardization, we were able to include all patients in the discharge population for accidental puncture or laceration and more than 95 percent of those for postoperative sepsis and central venous catheter-related bloodstream infection. After stratification, patient characteristics were identical between different types of hospitals. The excluded patients for PDIs represented sicker patients, with higher risk of experiencing postoperative sepsis and central venous catheter-related bloodstream infection. As a result of dropping these unmatched cases, the rates for PDI 10 and PDI 12 dropped substantially for large and small, teaching and nonteaching hospitals, though the effect was largest for teaching hospitals.

Teaching hospitals had higher indirectly standardized rates than nonteaching hospitals for all PDIs examined. However, the differences between teaching and nonteaching hospitals in directly standardized rates for accidental puncture or laceration (PDI 01) and postoperative sepsis (PDI 10) were not statistically significant.

Large hospitals had higher indirectly standardized rates of central venous catheter-related bloodstream infection (PDI 12). The differences in rates of accidental puncture and laceration and postoperative sepsis by bed size were not statistically significant. These results were consistent with the results for directly standardized rates by bed size.

Table IV.3. Observed, indirectly standardized, and directly standardized PDI rates, by hospital type (per 1,000 discharges)

		Teaching status			Bed size		
		Teaching	Nonteaching	Difference [95% CI]	Large[a]	Small[a]	Difference [95% CI]
PDI 01	Observed	0.83	0.23	0.60	0.72	0.34	0.39
	Indirect	0.63	0.50	0.12* [0.06, 0.19]	0.61	0.58	0.03 [-0.03, 0.10]
	Direct	0.63	0.59	0.04# [-0.03, 0.11]	0.61	0.57	0.04 [-0.04, 0.12]
PDI 10	Observed	21.33	11.25	10.08	21.29	15.69	5.60
	Indirect	20.99	16.17	4.82* [1.21, 8.43]	21.0	18.5	2.48 [-0.47, 5.43]
	Direct	15.23	15.14	0.08# [-4.75, 4.92]	18.36	16.39	1.97 [-1.72, 5.66]
PDI 12	Observed	1.75	0.27	1.48	1.48	0.54	0.95
	Indirect	1.31	0.62	0.70* [0.60, 0.79]	1.26	0.92	0.34* [0.23, 0.44]
	Direct	0.85	0.50	0.34* [0.23, 0.46]	0.89	0.61	0.28* [0.16, 0.41]

Sources: Mathematica Policy Research analysis of 12 SID from January 2009 to December 2010. FY 2010 AHA Survey Database.

Note: 95 percent CIs apply only to standardized rates. Teaching hospitals are defined as major or minor teaching hospitals in the 2010 AHA Survey Database.

[a] Large hospitals are those in the quartile with the largest number of beds (more than 271). Small hospitals are those in the quartile with the smallest number of beds (less than 56).

* Denotes that the difference in rates by hospital type is statistically significant at the 5 percent level.

# Denotes that comparing directly standardized rates by hospital type leads to a different conclusion regarding whether the differences are statistically significant.

PDI 01 = accidental puncture or laceration; PDI 10 = postoperative sepsis; PDI 12 = central venous catheter-related blood stream infection.

D. Recommendations

Our findings indicate that, generally, risk adjustment through indirect standardization adequately adjusts for different observed case mixes of patients between hospital types using rates directly standardized to the combined reference population as the standard for comparison. However, we determined that the differences in QI rates by hospital types changed with a direct standardization approach for some combinations of QIs and hospital characteristics. In particular, for one of the PSIs and two PDIs tested, approximately five to six percent of discharges were not represented in both hospital types for each comparison pair. These cases might be addressed by modifying QI denominator definitions to exclude cases that are not treated across hospital types. However, these cases appear to be important for differentiating among large and teaching hospitals. Disparities might also be addressed by stratifying results according to hospital type and only comparing results of hospitals of a particular type, or including additional terms in the logistic regression model that could better account for differences by hospital type not captured in the current models. We recommend investigating stratification into hospital or patient peer groups or adding risk factors to predictive models to address these cases. In addition, although we analyzed directly standardized rates to identify deficiencies in the current specification for indirectly standardized comparisons across hospital types, direct standardization could also be an appropriate primary approach for comparisons of QI rates for certain end uses/users.

Target audiences

For the purpose of making comparisons among many different hospitals, such as those in a large national population with a range of characteristics, indirect standardization is an appropriate method. The approach establishes the benchmark against which each hospital is compared, that is, expected performance for an average patient case mix in the reference population.

For assessment of hospital performance in smaller samples, such as hospital networks, risk adjustment through direct standardization methods could offer benefits over indirect standardization. The approach could be appropriate for any user aiming to assess performance for a specific population rather than the average population nationwide. For example, a hospital system might wish to evaluate the performance of its several hospitals against one another for quality improvement efforts. In this case, it would not be appropriate to formulate a benchmark through a hypothetical “average” hospital; instead, inferences about performance could be targeted by comparing the performance of each hospital against others in its treatment of a case mix of patients specific to that system. For example, Silber et al. (2014) suggest a template matching method that selects patient samples that are the same, on average, across multiple hospitals. If AHRQ is interested in expanding the functionality of the QI software to enhance performance assessment for internal quality improvement, then template matching or other direct standardization methods, could better serve users for this purpose.

Considerations for implementation

To perform risk adjustment through direct standardization, the software would require risk profiles defined based on the distribution of discharge-level risk factors within a reference population. Reference populations may be defined across the national population or by restriction to hospitals or patients with certain characteristics. For each hospital, the software would calculate QI rates within cells defined by the risk profiles. The software would then enable the user to calculate directly standardized rates and compare hospitals' rates for any of these risk profiles. For example, the software might include prevalences for combinations of risk factors for the national population, for the population treated by teaching hospitals, and for the population treated by nonteaching hospitals. It would also include each hospital's rate for each combination of factors. The combination of prevalence weights and rates would enable a hospital's directly standardized rates to be compared nationally and within the hospital's peer group. Rates adjusted using the teaching hospital weights would be suitable for teaching hospitals or for hospitals attempting to fill a role similar in some dimension to teaching hospitals in their local community. For example a nonteaching hospital that treats complex cases or that adopts advanced technology might fit in that peer group. To facilitate the types of comparisons that Silber et al. (2014) describe, users of the software could enter discharge-level data from two or more hospitals. Comparisons akin to those made in our study could be produced by the software.

Remaining unknowns

In our assessment of risk-adjustment methods, we assumed no relevant discharge-level risk factors were omitted from the risk-adjustment models that would contribute to differences in risk-adjusted rates by hospital type. For example, the difference in risk-adjusted pneumonia mortality rates (IQI 20) between teaching and nonteaching hospitals remains statistically significant, based on both indirect and direct standardization. If teaching hospitals are treating pneumonia patients with more complications that are not reflected in the data or not incorporated in the risk-adjustment models, neither method will be able to detect this. To address potential concerns about bias due to such omitted risk factors, we recommend that AHRQ conduct: (1) analyses to better understand whether such omitted risk factors exist and (2) sensitivity analyses to ascertain the magnitude of such bias that would produce the observed differences in risk-adjusted rates between hospital types. We discuss the potential concern related to such omitted (or unmeasured) risk factors in Chapter V.

Future analysis

For future study of the adequacy of AHRQ risk-adjustment models, the template matching method could be adopted to investigate variation in hospital QI rates at the level of individual hospitals. By this approach, AHRQ could assess the variation in hospital quality through risk-adjusted rates that are calculated on the same case mix of patients across hospitals. With tight control over variation in case mix, the remaining variation in QI rates by multiple hospital characteristics could be considered simultaneously, as well as their interactions (for example, large hospitals in urban and in rural areas). Additionally, building off current findings and the proposed future analyses, we propose that AHRQ also examine differences in QI rates by discharge risk profiles to ascertain the impact of heterogeneity on making inferences on hospital risk-adjusted rates, particularly across hospital types. We also recommend future study of other QIs that were not included in this analysis to determine the extent to which the conclusions drawn apply across all risk-adjusted QIs.

This page has been left blank for double-sided copying.

V. INCORPORATING HOSPITAL CHARACTERISTICS IN RISK-ADJUSTMENT MODELS

A. Methodological challenge

An ongoing debate among stakeholders involved in hospital quality measurement centers on the adequacy of risk-adjustment models used by hospital indicators (including the AHRQ QIs) . The debate has focused on whether differences in QI rates by hospital type are due to differences in unmeasured risk between hospital types that are unaccounted for in the current risk-adjustment models or to differences in quality of care (Medicare Payment Advisory Commission 2013, National Quality Forum 2014, Changes in Health Care Financing & Organization 2014).

The argument that differences are due to unmeasured risk postulates that certain hospital types treat higher proportions of patients with risk factors that are not reflected in the current risk-adjustment models (whether the relevant risk factors are unobserved, unmeasured, or just not included in the models). Hospitals treating higher proportions of patients with these risk factors than the norm would be at a disadvantage when hospitals are compared using the current risk-adjusted rates. For example, if teaching hospitals see higher proportions of patients with a risk factor for adverse events than nonteaching hospitals, these patients will experience higher rates of the adverse events (all other risk factors being equal) and teaching hospitals will have higher observed rates than nonteaching hospitals. However, the expected rates of teaching hospitals will not be correspondingly higher if the risk factor is not included in the risk-adjustment models, and risk-adjusted rates will be higher for teaching hospitals than nonteaching hospitals for that reason alone. Thus, comparisons between hospitals of different types could lead to spurious conclusions about the relative quality of care.

Conversely, the differences in rates for hospitals of different types could reflect differences in the average quality of care delivered by hospital type. In this case, the estimated rates accurately reflect the differences in performance across hospital types as intended, and no modification to the predictive model is needed. In fact, any modification to the method to account for additional risk factors could obscure differences in quality across hospital types.

In this chapter, we summarize: (1) how the modification could represent an improvement and a summary of the analytic approach, (2) the findings from the analysis, and (3) discuss recommendations that can be made in light of the remaining unknowns, including next steps to extend and build on the current analyses.

B. Potential improvement

We tested how incorporating an indicator of hospital type in the risk-adjustment models as an additional risk factor changed model fit and the results of comparisons between hospitals. We also tested whether estimating models separately for different hospital types changed the models substantially. If the differences in risk-adjusted rates documented in the EDA are due to unknown or unmeasured differences in patients that are correlated with risk of adverse events, including hospital characteristics in the risk-adjustment model could help account for these factors and further level comparisons of hospitals with different patient populations. Unfortunately, if differences in rates across hospitals reflect differences in quality across hospital types, these differences in quality would also be obscured by this adjustment, because average differences between hospital types incorporated in the models are adjusted to zero. Therefore, we are unable to say whether the changes to risk adjustment models that result from the addition of hospital type increase the accuracy of the QIs.

In the absence of definitive evidence regarding the roles that risk and quality play in causing the observed differences in hospital QI rates, we assessed whether incorporating hospital characteristics in the risk-adjustment models improved the accuracy of hospital comparisons while varying the proportion of the differences due to risk and quality. We used information from the actual relationships between hospital QI rates and hospital types to simulate hospital discharge data for which the differences in rates are due to a range of risk and quality mixes that we defined (all risk, all quality, and mixtures of both). We then tested how the modified models performed in ranking hospital quality compared to the current models. We discuss the overall analytic approach and the approach to the simulation analysis in more detail below.

Analytic approach

Our analytic approach was comprised of two components: (1) incorporating hospital characteristics in the current models to estimate the effect on QI results and (2) a simulation analysis that tested the potential improvement of the modified models. The point of comparison in all of the analyses described in the chapter was the current AHRQ risk-adjustment models estimated using the analytic file of hospital discharges in 12 states (base models). We added indicator variables for the hospital characteristics (for example, a variable indicating that a hospital is a teaching hospital) one at a time to the existing set of discharge-level risk factors in these base models. See Jones et al. (2014b) for a detailed description of the analytic approach to estimating the base and modified models.

We also explored an alternate way of incorporating hospital characteristics in the risk-adjustment models that is intended to account for potential correlation of hospital characteristics with discharge-level risk factors included in the models. The omission of this class of correlations in the risk-adjustment models could result in biased comparisons of hospital performance if the correlations are due to differences in unmeasured risk across hospital types. We account for these correlations by adding an average hospital type effect in the calculation of predicted values rather than a hospital type fixed effect. The modified models are estimated as described above with hospital type indicators added as risk factors (for example, an indicator for teaching hospitals), although instead of adjusting only hospitals of one type based on their hospital type effect (teaching hospitals), all hospitals receive an adjustment to their expected rate based on the average hospital type effect (average teaching and nonteaching effect) (Ash et al. 2011). Thus, the resulting expected rates are the predicted rates of adverse events for the average hospital type and with an average patient case mix; the latter is the case for all versions of the models. Although this alternate approach does not address the overall concern that there is unmeasured risk that varies by hospital type, it addresses a potential confounding issue with risk factors already included in the models and serves as a second point of comparison (with potentially lower bias) for the models that incorporate hospital type indicators in the models (the primary modification addressed in the analysis).

Simulation analysis. We used the 12-state discharge data to estimate key parameters forming the basis of the simulated data that are central to the analysis. The parameters are estimates of (1) the distribution of measured patient risk—estimated from the baseline models;

(2) the variance of risk-adjusted rates—estimated from the base models with hospital random effects; and (3) the differences in risk-adjusted rates by hospital type or hospital type effects—estimated from the base models plus an indicator of the specified hospital type, which is the primary model modification studied in the analysis.

The final step in simulating the patient data for the analysis was to vary the assumed mix of quality and risk (that is, vary the role that quality and risk play in causing the overall differences in rates by hospital type). We varied the mix of quality and risk under two simulated scenarios: (1) the proportion of the mean difference by hospital type and of the standard deviation of hospital rates due to risk vary together between 0 percent and 75 percent and (2) the proportion of the mean difference by hospital type due to risk varies between 0 percent and 100 percent, while the proportion of hospital standard deviation due to risk and quality are assumed the same.[13] These scenarios represent two different views of the relationship between quality, hospital characteristics, and differences in risk-adjusted rates by hospital type. The first scenario implies that the role of quality in differences by hospital type is proportionate to the role of quality in the variation in individual hospitals' QI rates; that is, the percentage of differences in the mean and spread of the distribution due to risk are equal. The second implies that the role of quality in the variation of hospital QI rates is independent of the mean difference by hospital type.

Next, we estimated hospital risk-adjusted rates using the simulated data and three versions of the risk-adjusted rates, which were produced by the base and modified versions of the risk-adjustment models:

- RAR(1): risk-adjusted rate generated from the base model, using the standard observed to expected ratio
- RAR(2): risk-adjusted rate generated from a model with a hospital type indicator, incorporating the same average hospital effect in the calculation of each hospital's expected rate
- RAR(3): risk-adjusted rate generated from a model with a hospital type indicator, incorporating the specific hospital type effect in the calculation of expected rates

In the final step of the simulation analysis, we tested the performance of the modification against the current risk-adjustment models in ranking hospitals according to simulated quality. We compared the correlation of the hospital's rank based on the three risk-adjusted rates to its rank based on simulated quality while also varying the proportion of the differences in the mean and variance of the rates due to quality versus risk. We compared the rank correlations across all hospitals and within hospital types.

We performed the simulations for two QIs, PSI 12 (postoperative deep vein thrombosis or pulmonary embolism) and IQI 20 (pneumonia mortality), and two hospital characteristics (teaching status and hospital bed size). We included only a subset of the QIs and hospital characteristics examined in this report due to computational limitations imposed by the simulation methods. We also simulated only two categories of bed size (the bottom two quartiles

[13] The proportion in scenario 1 cannot be 100, otherwise there would be no difference in quality to detect.

compared to the top two) to simplify the analysis. We chose one patient safety indicator and one mortality indicator to address the two major categories of adverse events indicated by the QIs. In addition, we chose these two QIs to examine the estimated effects for a range of values of the strength of relationships between hospital types and outcomes. In particular, we chose PSI 12 because inclusion of these two hospital characteristics to the base models produced the largest effect on model performance and hospital-level results compared to other QIs examined in this report (discussed earlier in the Results section). IQI 20 serves as a contrast to PSI 12, in that inclusion of teaching status in the risk-adjustment models appears to have a small effect on model performance and hospital-level rates and bed size only a somewhat larger effect. Also, IQI 20 is a commonly studied mortality indicator that is used in a variety of public and private programs, which could facilitate links to relevant findings in the literature. For more information on the simulation methods, see Jones et al. (2014b).

C. Findings

Impact of incorporating hospital characteristics. For most of the QI/hospital characteristics combinations, the magnitude of the relationships were small, and adding hospital type indicators did little to improve the fit and performance of the risk-adjustment models. Consequently, the changes in the mean rates, distributions, and classification of hospital by various methods were modest in these cases. For example, for QIs such as IQI 20 (pneumonia mortality), which demonstrated smaller associations with hospital characteristics, the changes in hospital rates and profiling by hospital type were relatively minor. However, there were exceptions in which the estimated relationships and resulting changes in hospital results were fairly sizable. For example, PSI 12 (postoperative PE/DVT) exhibited a strong association with teaching status after accounting for the discharge-level risk factor information; teaching hospitals were more likely to experience an event. Thus, adding an indicator for teaching status led to: (1) large changes in mean QI rates and the distributions of QI rates by hospital characteristics—in this case, a shrinking of the gap between teaching and nonteaching hospitals to zero and (2) large changes in the classification of hospitals—in this case, a shift in classification of teaching hospitals to better categories (lower rates) regardless of the profiling approach.

The hospital QI rates and rankings generated using the alternate approach of estimating the models separately by hospital type are nearly identical to those for the primary modified approach of adding hospital type indicators. We find that the relationships between the risk factors and adverse events varied for many of the risk factors when estimating separate models. However, our results did not suggest important differences in fit or hospital rankings when type was interacted with other risk factors compared to type entered as a simple fixed effect. In addition, we find that the hospital rankings for the alternate model incorporating an average effect are nearly identical to the current or base models, which is not surprising because most of the estimated average effects are quite small and the size of the effect applied to predicted values is the same for all discharges and hospitals.

To the extent that the differences in outcomes by hospital type are due to unmeasured risk, the resulting changes in hospital results from modifying the risk-adjustment models could represent an improvement in hospital comparisons. Conversely, the differences could be an indication of exactly what the QIs are intended to measure—differences in hospital quality indicated by their performance on the QIs. Because there is no strong evidence of the contributions of factors driving the differences, the magnitudes of the associations and changes in hospital results do not provide any evidence regarding whether the hospital characteristics improve the accuracy of the estimates; they simply provide a better understanding of the extent of the changes if the modification is made.

Simulation analysis. In all but one instance, the inclusion of hospital characteristics as risk factors obscured the quality signal contained in the risk-adjusted rate. Table V.1 shows how the proportion of hospital variation due to unexplained risk affects the correlation between quality and the different risk-adjusted rates.[14] Risk-adjusted rates containing hospital type risk factors [RAR(3)] showed reduced correlation of the risk-adjusted rate with hospital quality compared to the base model [RAR(1)] for IQI 20 and PSI 12. When we included an average hospital type effect to the calculation of hospital rates to reduce specification bias [RAR(2)], in no case did the inclusion make an appreciable difference in hospital rankings. Furthermore, although the addition of either teaching status or bed size to the models reduced the ability to rank all hospitals according to simulated quality, it had little effect on the models' ability to rank hospitals within hospital types; that is, within hospital type, the correlations of risk-adjusted rates with simulated quality were nearly identical in models with and without the hospital type indicators (results not shown).

Table V.1. Correlation of hospital quality and estimated rates under different mixes of quality and risk and analytic approaches: PSI 12 (postoperative PE or DVT) and IQI 20 (pneumonia mortality), by teaching effects

		IQI 20			PSI 12		
	Variation due to risk	RAR(1)	RAR(2)	RAR(3)	RAR(1)	RAR(2)	RAR(3)
Mean and spread	0%	0.785	0.785	0.784	0.702	0.702	0.617
	25%	0.653	0.653	0.649	0.615	0.615	0.534
	75%	0.194	0.194	0.193	0.239	0.239	0.129
Mean only	0%	0.499	0.499	0.497	0.508	0.508	0.318
	25%	0.485	0.485	0.483	0.480	0.480	0.346
	75%	0.425	0.425	0.425	0.395	0.395	0.340
	100%	0.440	0.440	0.440	0.349	0.349	0.381*

Sources: Mathematica Policy Research analysis of SID data from 12 states from January to December 2010. Teaching status determined from FY 2010 AHA Survey Database.

Note: RAR(1) refers to the risk-adjusted rates from the base model. RAR(2) refers to the risk-adjusted rates from the model with an indicator for teaching status added and expected rates calculated using the average hospital type effect. RAR(3) refers to the risk-adjusted rates from the model with an indicator for teaching status added and expected rates calculated using the hospital type effect.

PE = pulmonary embolism; DVT = deep vein thrombosis.

* Denotes the only instance that incorporating an indicator for teaching status as a risk factor to the risk-adjustment model resulted in an increase in the correlation with quality, that is, an improvement in the ability to rank hospitals according to quality.

[14] As expected, when the proportion of hospital variation due to risk increases, the correlation between rankings based on quality and rankings based on the risk-adjusted rate decreases (reading from top to bottom in a given column). The result is true no matter which risk adjusted-rate is used, because the additional random variation in patient outcomes due to risk obscures the quality signal that differentiates hospitals.

For IQI 20, because there is little difference by hospital type, the degree of correlation is not noticeably affected by the risk-adjustment method. However, for PSI 12, adding teaching status as a risk factor [RAR(3)] reduces the correlation, with the direction and size of the effect depending on the way the mean difference by type and the spread of the distribution are allocated between risk and quality.[15]

The exception to this finding occurred when the difference in hospital type mean effects was assumed to be entirely due to risk, but hospitals' variation in quality and hospital risk were assumed similar in magnitude. As the role of risk in the mean difference by type increases, the benefit of controlling for hospital type in the risk-adjusted rates increases until it becomes greater than the cost of obscuring hospital-level quality variation. All or virtually all of the mean difference by hospital type must be attributed to risk before including type as a risk factor increases the ability of the risk adjusted rate to detect quality. Reducing the assumed role of risk to 75 percent of the mean difference reversed any improvement in identifying quality. The analysis of the relationships between PSI 12 and IQI 20 and bed size yielded results that were consistent with the findings presented for the two QIs and teaching status (results not shown).

D. Recommendations

Based on the evidence provided in the simulation analysis, we recommend that hospital type indictors not be incorporated as risk factors in the risk-adjustment models. The inclusion of hospital characteristics in the models in this way led to an improvement in distinguishing hospital quality under very limited circumstances (that is, when all of simulated quality was due to unmeasured risk) while obscuring quality in all other circumstances. Furthermore, when adding hospital characteristics to the models as a correlate of quality (that is, incorporating average hospital type effects as an alternate approach), there is very little change in the rankings of hospitals by their QI rates. Combined with the findings in the analysis of standardization approach, these findings suggest that alternate approaches to presenting results for multiple types of hospitals that consider hospital types in the comparisons (such as stratification/peer grouping or direct standardization approaches) might be desirable when there are large differences in QI rates by hospital type.

Target audiences

The evidence from the simulation analysis suggests that including hospital type indicators as risk factors is not advisable for any use that compares QI rates across hospital types when it is not certain that all or nearly all of the observed differences in QI rates are due to differences in unmeasured risk. Adding hospital type indictors to the models as a risk factor removes any average differences in risk-adjusted rates by hospital type. If an average hospital type has a higher rate because average quality is lower, every hospital of that type will be adjusted downward (better) under this approach. This adjustment will obscure a quality signal embedded in the rates of these hospitals. Before conducting the simulation analysis, we knew that this type

[15] When hospital-level variation including the mean difference by type is attributed primarily to risk (75 percent of variation in mean and spread due to risk), including hospital type in risk adjustment results in the largest decreased in percentage terms of the correlation with simulated quality. In that case, the weak quality signal is swamped by the estimated hospital type effect. We do not consider cases in which 100 percent of variation is due to risk, because such cases do not have quality variation to detect.

of adjustment would occur to some extent (and likely mean that comparisons across hospital types are not advisable), but there was the potential benefit of accounting for unmeasured risk in improving the accuracy of estimated hospital risk-adjusted rates. However, this potential benefit did not offset the obscuring of quality except in the extreme case in which all or nearly all of the differences were due to risk. In addition, for uses that do not compare rates across hospital types, the modification is of little benefit. For comparisons within a hospital type, if the primary objective is to compare hospital rates to one another, there is no benefit of adjusting hospital rates by a single factor for all hospitals in the group (that is, the net effect on relative ranking will be zero).[16]

If users of QI results are concerned that comparisons between hospitals of different types are distorted by differences in unmeasured risk, they might adopt a more restricted method of comparison rather than altering the risk-adjustment methods. For example, a user could stratify or peer-group hospital types and compare the rates within a type (for example, teaching hospitals are compared to teaching hospitals and nonteaching hospitals to nonteaching hospitals). This information could be coupled with comparisons across all hospitals to provide hospitals and policymakers with a hospital's performance relative to national and peer-group benchmarks. Stratification by hospital type is equivalent to incorporating a hospital type risk factor in the model, except that it prevents comparisons across hospital types. Preventing such comparisons may be important in contexts such as public reporting. In other contexts, however, the approach of restricting comparisons may be completely equivalent in effect to incorporating hospital type as a risk factor. For example, if a program rewards hospitals a bonus based on their risk-adjusted performance relative to their peer group benchmark, the distribution of rewards may be the same as the distribution resulting from risk adjustment that includes hospital type.

Considerations for implementation

Any approach to hospital comparisons that incorporates information on hospital types must determine how to define and group hospital types. The challenge is further complicated if there is reason to consider multiple hospital characteristics in defining the hospital types (for example, combining hospital size and teaching status to define hospital groups). We recommend starting any discussion on this topic with an assessment of the conceptual rationale for why rates differ by hospital type. Once the rationale is established, empirical evidence regarding how various hospital groups differ by QI rates can be used to provide support for defining the groups. Ultimately, the decisions will also largely depend on the data available to classify hospitals into groups.

Remaining unknowns

Because we begin from a position of less than full information regarding the roles of quality and unmeasured risk in the hospital type effects we measure, the size of those effects tells us nothing about how or whether they should change the way we estimate hospital rates. Simulating the contributions of these factors helps us to think about which modifications are likely to lead to

[16] If the accuracy of each hospital's rate is paramount (rather than the relative rates across hospitals), applying the hospital type adjustment could lead to an improvement (Note: Comparison of hospital rates, which is the focus of this project, is not the objective in this case). However, once again, if a portion of the differences in rates by hospital type is due to differences in quality of care, the modification could reduce the accuracy of rates for this purpose.

improvements in the accuracy of hospital comparisons. The findings indicate clearly that including hospital type in the risk adjustment model itself is likely to be helpful only in the extreme case that risk is the only factor contributing to the differences. However, tests of other potential improvement are not likely to lead to such clear recommendations. Thus, definitive recommendations on modifications to the results for specific end uses/users will need to be informed by increased evidence regarding the observed differences.

In addition, this simulation analysis is subject to several limitations. Besides the limitation that hospital characteristics and known risk factors are assumed to be independent in the modified models, several philosophic and technical limitations also exist. The assumption of a normal distribution for both quality and risk, although made for analytic tractability, is not well founded and, in fact, could be far from reality for some QIs given the skewed distributions of observed risk and hospital rates. Similarly, only a handful of variants on the quality and risk distribution were tested, and the results for these limited cases suggest that the point at which a characteristic should be considered as a risk factor is when it is clear that quality has no role in the difference. Logical extensions of the simulation analysis include considering the correlation between risk factors and hospital characteristics, a wider range of assumptions concerning quality and risk distributions, and also a systematic search for the point at which the treatment of a characteristic as a risk factor improves the performance of the models.

Further analysis

Including hospital characteristics in risk-adjustment models as we did in our test specifications is inadvisable because including hospital type indicators is a blunt instrument for the purpose of accounting for differences in the unmeasured risk of hospital patient populations. The modification adjusts all hospitals of a type to the same extent and applies no adjustment to their counterparts, regardless of their patient case mix and risk. To the extent that unmeasured risk varies greatly within a hospital type (for example, teaching hospitals may have greater aggregate unmeasured risk on average, but it will likely vary substantially across teaching hospitals), this modification will oversimplify the relationship by applying the same adjustment to all hospital of the type. The simulation analysis provides evidence that the approach does more on average to obscure quality than it does to capture a hospital's unmeasured risk.

We recommend examining the inclusion of discharge-level variables (such as measures of clinical, sociodemographic, or socioeconomic characteristics) that could proxy for patient risk in the risk-adjustment models as a potential solution to hypothesized differences in unmeasured risk by hospital type. Discharge-level variables show greater promise in functioning as proxies of unmeasured risk than hospital-level characteristics because the adjustment is capable of varying for each hospital depending on its patient case mix rather than a one-size-fits-all adjustment based on the hospital type. For this reason, adding discharge-level variables as proxies for risk allows for potential improvement (accounting for unmeasured risk that varies by hospital type on average) without necessarily obscuring differences in quality by hospital type (as long as patients with the given discharge-level factor do not receive lower quality of care on average). We recommend extending the current analysis to test whether the hospital type effects we have identified are associated with known patient characteristics, such as clinical, sociodemographic, or socioeconomic characteristics, and test the effect of their inclusion in the risk-adjustment models. However, if patient characteristics are associated with differences in the quality of care, including them in risk adjustment will obscure variations in quality of care just as including hospital characteristics will. Thus, we also recommend further analysis to explain differences in QI rates associated with these patient characteristics. We motivate and summarize potential future analyses on these topics in Chapter XIII, Discussion.

This page has been left blank for double-sided copying.

VI. SHRINKING TO ALTERNATE TARGETS

A. Methodological challenge

The accuracy of hospital reliability-adjusted rates relies on informed, evidence-based prior distributions (shrinkage parameters) used in the reliability-adjustment process; however, for certain applications of the QIs, shrinkage parameters other than those estimated using the HCUP reference population provided with the AHRQ QI software may produce more valid reliability-adjusted rate estimates.[17] The potential benefits of choosing an alternative prior largely depends on the analytic sample and how it compares to the reference population.[18] If the analytic sample is a random subset of the HCUP reference population, then the default shrinkage parameters are likely appropriate, assuming that the default prior assumption of no variation in rates by hospital type is correct. However, if the analytic sample is a select set of discharges (for example, Medicare patients) or providers (for example, teaching hospitals), then it could be desirable to reestimate the shrinkage parameters, which would use a more fitting set of discharges in calculating hospitals' rates. The approach of using parameters estimated across all patients in the HCUP reference population could produce shrinkage parameters that are inconsistent with the analytic samples, limiting the accuracy of the estimated rates and hospital comparisons. Similarly, when the analytic sample is external to the reference population because it covers a different time period, or the analytic sample uses a different data source, using the parameters from the HCUP reference population may limit the accuracy of the results. In addition, when comparing hospital rates across hospital types, it is possible that more information is available on the quality of different hospital types to be used in the reliability-adjustment process (that is, more appropriate priors). By not considering hospital type in the prior, the current approach could be ignoring information on differences in quality by hospital type. As such, reestimating the shrinkage parameters for certain analytic samples or by hospital type could improve the accuracy of estimated hospital rates and resulting comparisons using the rates.

B. Potential improvement

We examined three modifications to the method of estimating shrinkage parameters used for the AHRQ QIs. Instead of using the parameters estimated on the overall reference population, the modifications use parameters estimated from the analytic sample or estimated separately by peer group. The objective is to estimate shrinkage parameters (mean and variance) that are better representations of true hospital quality for the population to be analyzed than provided currently using the reference population, leading to more accurate comparisons of quality across hospital types. The first two modifications to the approach are applicable when the analytic sample is substantially different than the reference population in some way. The third modification applies if there are differences in hospital quality by hospital type.

[17] Even if risk-adjusted rates are unbiased for hospital types, incorporating additional information in the prior can reduce random error in the reliability-adjusted rates by shrinking hospital estimates to rates that are more likely to reflect their "true" rates.

[18] For the purposes of this analysis, the analytic sample refers to the set of discharges and hospitals over which a user aims to estimate hospital rates.

- **Reestimating the shrinkage parameters (partial recalibration)**. There are two options when reestimating the shrinkage parameters using the analytic sample: reestimating only the mean or reestimating the mean and the signal variance.
- **Reestimating the shrinkage parameters and the risk-adjustment models (full recalibration)**. The benefit of this approach is that it aligns the risk- and reliability-adjustment process to the analytic sample, making for a more consistent and comprehensive approach across the risk- and reliability-adjustment processes and increasing the interpretability of results across risk- and reliability-adjusted rates. Even though the analysis of risk adjustment described above suggests that reestimation will produce little effect on parameter estimates, the benefit is to make consistent all of the parameters used to produce estimates of hospital QI rates. Analytic samples with large numbers of discharges are required to reestimate the risk-adjustment models.
- **Reestimating the shrinkage parameters separately by peer group (peer group shrinking)**. The first step is to reestimate the mean for each peer group; thus, hospital risk-adjusted rates are shrunken to their hospital type (peer group) mean rather than the national mean from the HCUP reference population. The rationale for this approach is that the peer group means could contain more information on hospital quality for use in reliability adjustment than does the reference population mean alone. However, there are potential drawbacks of this approach: (1) If rates differ by hospital type due to some factor other than quality, shrinking to hospital type targets exacerbates these differences; and (2) The approach makes a strong statement about difference in quality by hospital type, which will not necessarily be true for all hospitals in a peer group, just on average. A user can also reestimate signal variance separately by hospital type if there is a conceptual rationale for doing so and the dataset is large enough.

Analytic approach

We ran the software four ways to support our analyses using the HCUP and Medicare datasets. First, we ran the software as published to produce reliability-adjusted rates using the published risk-adjustment coefficients and published shrinkage parameters based on the HCUP reference population (published model), representing a baseline comparator for all analyses. Second, we ran the software using the published risk-adjustment coefficients but reestimated the shrinkage parameters on the analytic sample (partial recalibration). There are two options for partial recalibration: calculating the mean only, or reestimating the mean and signal variance. Third, we ran the software reestimating the risk-adjustment models and shrinkage parameters on the analytic sample (full recalibration). Fourth, using the fully recalibrated risk-adjusted rates, we reestimated the shrinkage parameters for select peer groups (peer group shrinking). For the peer group shrinkage approach, we used the reestimated risk-adjustment models in order to consider the peer groups as select subsets.

We estimated the effect of the three modified approaches on reliability-adjusted rates, reliability weights, and performance categories using two datasets representing a select subsample (that is, a subset of the HCUP reference population to illustrate the effects on a sample with characteristics similar to the reference population) and an external sample (that is, a completely different data source external to the reference population to illustrate the effect on a population that is likely quite different than the reference population).

- The select subsample is the 12-state 2010 HCUP SID described in Chapter II. We consider the 12 states a select subsample because version 4.5 of the QI software uses the HCUP SID from 44 states from 2010 as the reference population.
- The external sample contains all Medicare fee-for-service discharges at inpatient prospective payment system (IPPS) hospitals from April 2011 through March 2012. The Medicare fee-for-service sample includes hospitals from all states and the District of Columbia, less Maryland hospitals (all of which are exempt from IPPS rules).

For each dataset, we calculated PSI, IQI, and PDI rates using SAS version 4.5 of the QI software. For this analysis, we focused on PSIs 6, 12–15; IQIs 15, 16, and 20; and PDIs 1, 10, and 12. This subset covers a range of clinical properties (that is, a range of adverse events covered by the QIs) and statistical characteristics (for example, relatively rare events as well as the most common events covered by the QIs). We also prioritized QIs that have the largest weights in the calculations of hospital composite values. Moreover, many of these QIs demonstrated empirical relationships with hospital characteristics examined in the EDA (Jones et al. 2014a).

When comparing the alternative approaches to reliability adjustment, we reviewed their effects overall and by hospital type on: (1) shrinkage parameters (mean and signal variance), (2) average reliability weights, and (3) hospital performance. Changes to the shrinkage parameters and reliability weights provide information regarding how the distribution of reliability-adjusted rates will likely change for each modification. Thus, we report whether a parameter increases or decreases when it is reestimated. The direction of change in the shrinkage target will be the same as that of hospitals' reliability-adjusted rates. The signal variance is a key factor in determining the spread of the distribution of reliability-adjusted rates. An increase in signal variance will increase the spread of the distribution. Greater reliability, measured by the reliability weight (a ratio of the signal variance to a hospital's noise or within variance), also translates to greater spread in the reliability-adjusted rate distribution.

We use three methods to test the effects of reestimating the reliability-adjustment parameters on hospital performance: changes to hospital performance categories assigned by comparing hospitals' reliability-adjusted rates to the national mean; the Kolmogorov-Smirnov (KS) test for differences in the distribution of rates; and the Spearman correlation of hospital ranks. Performance categories result from testing the hypothesis that hospitals' reliability-adjusted rates are equal to the national mean and are defined by comparing the 95 percent CI around a hospital's reliability-adjusted rate against the reference population rate for a given QI. When the upper bound of the CI is less than the national average, a hospital is classified as better than average. Hospitals with lower bounds that are higher than the national average are classified as worse than average. All other hospitals are deemed no different from the average.[19] To understand whether the change in shrinkage target or signal has a meaningful impact on hospital comparisons, we report the proportion of hospitals that move across performance categories.

[19] Note that this performance categorization method is a frequentist method that assumes hospital estimates are normally distributed, not an empirical Bayes method. We discuss the distributional assumptions of the framework in more detail in Chapter VII.

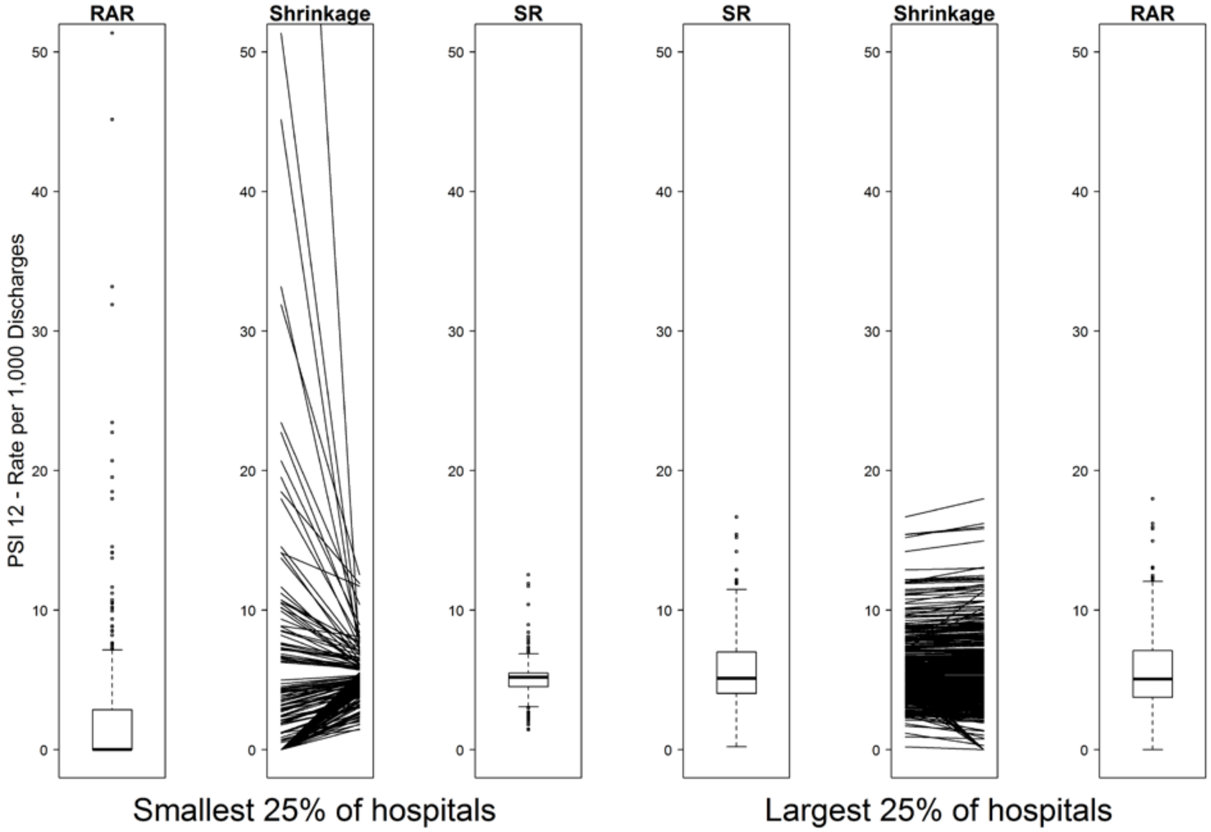
In addition , because reestimating the shrinkage parameters and risk-adjustment models could have a substantial effect on the rank of hospitals by their reliability-adjusted rates (whereas reestimating just the shrinkage parameters will primarily affect the spread and center of the distributions), we performed nonparametric tests of the effect of recalibration on the distribution of rates. We tested for differences in the reliability-adjusted rate distribution using the KS test. In addition, we tested for a difference in reliability-adjusted rate rankings using the Spearman rank-sum correlation test. For the approaches where shrinkage parameters are fit separately by peer group, we bootstrapped the calculation of the shrinkage targets and signal variance to test the stability of these estimates because sample sizes are small for some peer groups. In addition, we compared the mean and signal variance for the hospital types to determine whether the differences are statistically different.

C. Findings

Although the effect of the shrinkage approach varies depending on the QI and the analytic sample, three key findings related to hospital size can be applied across the three modifications (reestimate shrinkage parameters, shrinkage parameters and risk-adjustment models, and separately by hospital type) and the two types of analytic samples (select subsample and external sample). We summarize these key overarching findings below, and then summarize additional findings reported by the two types of analytic samples and the modification that reestimates shrinkage parameters separately by hospital type.

1. **Small hospital reliability-adjusted rates are the most sensitive to a change in the shrinkage target.** Using postoperative PE/DVT (PSI 12) as an example, Figure VI.1 shows the distance (shrinkage) that risk-adjusted rates (represented by the risk-adjusted rate [RAR] box plot) are pulled toward the shrinkage target (the mean of the reliability-adjusted rate distribution). Because they have the lowest reliability weights, small hospitals are shrunken close to the shrinkage target, as shown by the sloping lines. On the other hand, the parallel shrinkage lines show that the risk-adjusted rates and reliability-adjusted rates for large hospitals are strongly correlated. In general, the smallest hospitals are always pulled closer to the shrinkage target, irrespective of the shrinkage approach.

Figure VI.1. Extent of shrinkage, by hospital size



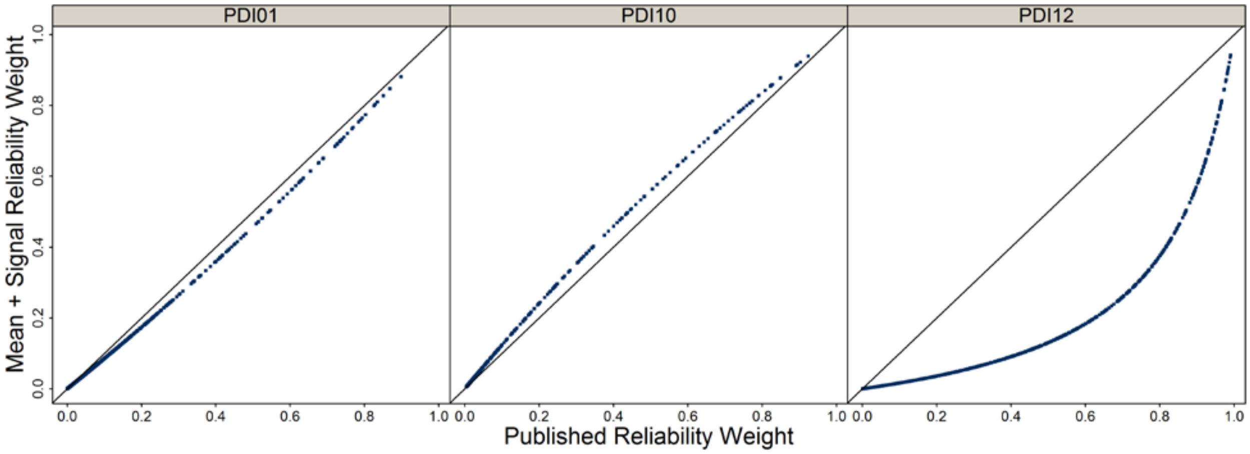
Sources: Mathematica analysis of SID data from 12 states from January 2009 to December 2010 using AHRQ software version 4.5; FY 2010 AHA Survey Database.

Note: The smallest hospitals are those in the first quartile for total number of licensed beds. The largest hospitals are those in the fourth quartile for total number of licensed beds.

RAR = risk-adjusted rate; SR = shrunken or reliability-adjusted rate.

2. **Reestimating signal variance has the greatest impact on the reliability weights for medium-size hospitals.** Each hospital's reliability weight is calculated as the ratio of the signal variance to the total variance (signal plus noise variance). Figure VI.2 shows changes in reliability weights for pediatric accidental puncture of laceration, postoperative sepsis, and central venous catheter-related blood stream infection (PDIs 01, 10, and 12) when reestimating the signal variance. The dotted line shows that the greatest changes in reliability are concentrated near a reliability weight of 0.5 for the published model, which are often medium-size hospitals. This finding is due to the fact that, when signal variance changes, the numerator and denominator of the reliability weight changes. Holding noise variance constant, hospitals with reliability weights close to 0.5 are most sensitive to these changes.

Figure VI.2. Effect of the within-sample approach on PDIs 01, 10, and 12 reliability weights



Sources: Mathematica analysis of SID data from 12 states from January 2010 to December 2010 using AHRQ software version 4.5; bed size and teaching status from FY 2010 AHA Survey Database; disproportionate share hospitals from FY 2010 CMS Impact File.

Note: The mean + signal reliability weights are estimated from the models reestimating mean and signal variance (partial recalibration). The published reliability weights are estimated from the current AHRQ QI models.

PDI 01 = accidental puncture or laceration; PDI 10 = postoperative sepsis; PDI 12 = central venous catheter-related blood stream infection.

3. **Medium or large hospitals are most likely to change performance categories.** Small hospitals are almost always classified as no different from average because their reliability-adjusted rates are pulled toward the shrinkage target; this is true even when shrinking to peer group targets. Therefore, although the reliability-adjusted rates for small hospitals are affected the most by a change in the shrinkage target, the effects tend to be quite small on performance categorization because the confidence intervals for small hospitals' rates remain wide on average under the alternate approaches. The only hospitals with reliability-adjusted rates that are far from the shrinkage target are medium or large hospitals.

Findings for select subsample

When reestimating shrinkage parameters on the 12-state SID sample, the shrinkage targets, average reliability, and performance categories change slightly, and the magnitude and direction of the effect depend on the QI and the approach (Table VI.1). When only reestimating the mean, shrinkage targets increase by as much as 13 percent and decrease by as much as 7 percent. Shrinkage targets increase for the IQIs and PDIs, but the pattern is mixed for the PSIs. The average reliability does not change for each QI because signal and noise are unchanged. In the mean-only and published approaches, signal and noise are based on the reference population used by the software. Compared with the original models, the mean-only approach has a minimal effect on performance categories; less than one percent of hospitals shift between categories for all but one QI (postoperative sepsis, PDI 10).

Table VI.1. Changes in shrinkage targets and average reliability, by approach

	Comparing reestimating mean to original approaches			Comparing reestimating mean-and-signal to original approaches		
QI	Shrinkage target (ratio)	Average reliability (ratio)	Performance categories (percentage)	Shrinkage target (ratio)	Average reliability (ratio)	Performance categories (percentage)
PSI 06	0.96	NA	0.34	0.97	1.20	0.41
PSI 12	1.08	NA	0.74	1.00	0.99	0.08
PSI 13	1.08	NA	0.85	1.06	1.09	1.60
PSI 14	0.94	NA	0.00	0.91	0.49	0.09
PSI 15	0.93	NA	0.95	0.90	0.95	1.42
IQI 15	1.03	NA	0.88	1.02	1.05	1.25
IQI 16	1.02	NA	0.62	1.04	1.08	5.14
IQI 20	1.05	NA	0.56	1.06	1.06	4.67
PDI 01	1.03	NA	0.00	1.06	0.97	0.09
PDI 10	1.12	NA	1.40	1.14	1.03	2.10
PDI 12	1.13	NA	0.00	1.14	0.82	1.00

Sources: Mathematica analysis of SID data from 12 states from January 2010 to December 2010 using AHRQ software version 4.5.

Note: For the mean-only approach, the shrinkage target is the overall observed-to-expected rate ratio. The shrinkage target for the mean-and-signal approach is the average of hospital risk-adjusted rates, weighted by the inverse variance squared. Average reliability is the average of hospital signal-to-noise ratio, weighted by the inverse variance squared. Performance categories are calculated by comparing the 95 percent CI around the reliability-adjusted rate to the reference population rate.

NA = Not applicable, reliability weights do not change when only reestimating the mean.

PSI 06 = iatrogenic pneumothorax; PSI 12 = postoperative pulmonary embolism or deep vein thrombosis; PSI 13 = postoperative sepsis; PSI 14 = postoperative wound dehiscence; PSI 15 = accidental puncture or laceration; IQI 11 = abdominal aortic aneurysm repair mortality; IQI 15 = acute myocardial infarction mortality; IQI 16 = heart failure mortality; IQI 20 = pneumonia mortality; PDI 01 = accidental puncture or laceration; PDI 10 = postoperative sepsis; PDI 12 = central venous catheter-related blood stream infection.

Table VI.1 also shows how reestimating the mean and the signal variance affect the reliability-adjusted rate distribution. The spread of the reliability-adjusted rate distribution is substantially different for several QIs when mean and signal variance are reestimated, but this difference does not necessarily translate into a change in performance category. For example, the average signal-to-noise ratio for postoperative wound dehiscence (PSI 14) decreased 51 percent, but less than one-tenth of one percent of hospitals moved performance categories. In the case of PSI 14, few hospitals moved because the average reliability for the QI is relatively low, and few hospitals are categorized as different from the mean. Conversely, the reliability-adjusted rate distributions for AMI, heart failure, and pneumonia mortality (IQI 15, 16, and 20) experienced moderate increases in reliability, which changed performance categories for 1 to 5 percent of hospitals.

Findings for external sample

The effect of alternative shrinkage approaches on results using the external sample was similar to the effect of the alternate approaches using the select sample (Table VI.2). Under the mean-only approach, shrinkage targets changed reliability-adjusted rates by at most 11 percent compared to the baseline results generated from the published model, but less than two percent of hospitals move performance categories. Reestimating the mean and the signal, as with other approaches, the change in the signal variance seems independent of changes in the shrinkage target. For example, the shrinkage target is unchanged for accidental puncture or laceration (PSI 15), but its signal variance decreases. Moreover, as with other approaches, large changes in shrinkage parameters do not necessarily lead to large changes in performance categories. We present results only for the PSIs for ease of exposition, but the high-level findings are representative of the IQIs and PDIs examined in this analysis.

Table VI.2. Changes in shrinkage targets, average reliability, and performance category, by shrinkage approach

Comparison	Statistic	PSI 06	PSI 12	PSI 13	PSI 14	PSI 15
Mean-only vs. original approach	Shrinkage target (ratio)	1.01	1.11	1	1.09	0.86
	Average reliability (ratio)	NA	NA	NA	NA	NA
	Performance categories (percentage)	0.00	0.95	0.00	0.09	1.62
Mean-and-signal vs. mean-only approach	Shrinkage target (ratio)	1.02	0.96	1.01	0.98	1
	Average reliability (ratio)	1.18	1.04	0.99	1.55	0.9
	Performance categories (percentage)	0.15	0.29	0.05	0.43	0.93

Sources: Mathematica analysis of Medicare fee-for-service claims from April 2011 to March 2012 using AHRQ software version 4.5; teaching status from FY 2010 AHA Survey Database.

Note: For the mean-only approach, the shrinkage target is the overall observed-to-expected rate ratio. The shrinkage target for the mean-and-signal and full recalibration approaches is the average of hospital risk-adjusted rates, weighted by the inverse variance squared. Average reliability is the average of hospital signal-to-noise ratio, weighted by the inverse variance squared. Performance categories are calculated by comparing the 95 percent CI around the reliability-adjusted rate to the reference population rate.

NA = Not applicable, reliability weights do not change when only reestimating the mean.

PSI 06 = iatrogenic pneumothorax; PSI 12 = postoperative pulmonary embolism or deep vein thrombosis; PSI 13 = postoperative sepsis; PSI 14 = postoperative wound dehiscence; PSI 15 = accidental puncture or laceration.

Full recalibration had a varied effect depending on the PSI (Table VI.3). Compared to the original approach published with the QI software, full recalibration changed the shrinkage target by more than 12 percent for all PSIs. The effect of full recalibration on reliability was greater than the mean-only or the mean and signal approach, leading to the largest observed changes in performance categories. Recalibration changed the shape of the reliability-adjusted rate distribution substantially for all PSIs (all KS tests significant at the 0.001 level), but the ranking of reliability-adjusted rates was nearly the same to the mean-and-signal approach (all Spearman correlations greater than 0.98). However, even though ranks were highly correlated, classifications changed substantially for two QIs. Full recalibration had the greatest impact on classification of postoperative PE/DVT and accidental puncture or laceration rates (PSIs 12 and 15, respectively), in which performance categories shifted for 5.6 and 8.5 percent of hospitals, respectively.

Table VI.3. The effect of reestimating risk-adjustment models and shrinkage parameters on PSI rates relative to the original approach

	Shrinkage target (ratio)	Average reliability (ratio)	Performance categories (percentage)	Spearman correlation	KS test
PSI 06	1.12	1.16	0.33	0.997	<0.001
PSI 12	1.35	1.00	5.62	0.991	<0.001
PSI 13	1.16	0.97	0.86	0.990	<0.001
PSI 14	1.51	1.20	0.87	0.996	<0.001
PSI 15	0.79	0.96	8.50	0.980	<0.001

Sources: Mathematica analysis of Medicare fee-for-service claims from April 2011 to March 2012 using AHRQ software version 4.5; teaching status from FY 2010 AHA Survey Database.

Note: All ratios compare the full recalibration to the original approach. The full recalibration approaches is the average of hospital risk-adjusted rates, weighted by the inverse variance squared. Average reliability is the average of hospital signal-to-noise ratio, weighted by the inverse variance squared. Performance categories are calculated by comparing the 95 percent CI around the reliability-adjusted rate to the reference population rate. Spearman correlations compare the rank of reliability-adjusted rates.

KS = Kolmogorov-Smirnov. PSI 06 = iatrogenic pneumothorax; PSI 12 = postoperative pulmonary embolism or deep vein thrombosis; PSI 13 = postoperative sepsis; PSI 14 = postoperative wound dehiscence; PSI 15 = accidental puncture or laceration.

Peer group results

To examine the effect of estimating parameters that differ by hospital peer group, we reestimated the shrinkage parameters separately for teaching and nonteaching hospitals. We focus on postoperative DVT/PE (PSI 12) as an example of the likely upper bound of effects because our previous analyses have demonstrated the largest difference in risk-adjusted rates for teaching and nonteaching hospitals for this QI. The analysis uses Medicare data with fully recalibrated risk- and reliability-models as the baseline comparator. Using this setup, the observed-to-expected rate ratio across all teaching hospitals is 45 percent greater than the ratio for nonteaching hospitals (results not shown).

Refitting the shrinkage target and signal variance by teaching status has a modest impact on PSI 12 results (Table VI.4). For the 25 percent of hospitals classified as teaching hospitals, the shrinkage target increased by 17.6 percent when using the peer-group approach relative to the baseline model; the signal variance increased 42.7 percent. In contrast, the shrinkage target for nonteaching hospitals decreased by 13.7 percent; the signal variance decreased by 50.0 percent. These changes indicate that the reliability-adjusted rates and the spread of the distribution increased for teaching hospitals; whereas the rates and spread of the distribution of rates decreased for nonteaching hospitals. Comparing the teaching and nonteaching reliability-adjusted rate distributions, the shape is significantly different (KS test p-value <0.01). These differences lead to changes in performance categories for 1.96 percent of teaching hospitals and 1.75 percent of nonteaching hospitals.

Table VI.4. Shrinkage parameters and change in performance category for postoperative PE/DVT (PSI 12), by teaching status

Statistic	All Hospitals (Baseline Recalibrated Model)	Teaching	Nonteaching
Number of hospitals	3,256	783	2,473
Shrinkage target	5.91	6.95	5.10
Signal variance	9.64	13.76	4.92
Performance category change (percentage)	NA	1.96	1.75

Sources: Mathematica analysis of Medicare fee-for-service claims from April 2011 to March 2012 using AHRQ software version 4.5; teaching status from FY 2010 AHA Survey Database.

Note: Shrinkage target and signal variance have units per 1,000 discharges. Hospital noise variance is derived from fully recalibrated Medicare results. Shrinking targets are the average of hospital risk-adjusted rates for each peer group, weighted by the inverse variance squared. Baseline model is the Medicare fee-for-service data with fully recalibrated risk- and reliability-adjustment models.

PSI 12 = postoperative pulmonary embolism or deep vein thrombosis (PE/DVT); NA = not applicable.

The effect of peer group shrinkage is highly dependent on the QI of interest and the specification of the risk-adjustment model for the QI, which is applied to estimate hospital noise variance. Using the Medicare analytic sample, performance categories for PSIs 06, 13, 14, and 15 shifted for 0 to 3 percent of hospitals when using the peer-group approach compared with the baseline approach (results not shown). Instead of refitting both shrinkage parameters by peer group, we could have only refit the shrinkage target and assumed equal variance. However, in addition to different shrinkage targets (t-test p-value <0.01), the differences in signal variance between teaching and nonteaching hospitals are statistically significant (F-test p-value <0.01). In the PSI 12 example, assuming unequal variances is warranted and feasible, because the number of hospitals in each peer group is large, and PSI 12 has relatively high reliability compared with other QIs. Our analysis of the stability of signal variance estimates shows that signal variance is increasingly uncertain as the peer group, number of discharges per hospital, or incidence of PSI event decreases.

D. Recommendations

Our findings provide evidence that the reestimation of shrinkage parameters can have a meaningful effect on reliability-adjusted rates, reliability weights, and performance categorization. That effect differs substantially depending on which parameters are reestimated, whether peer group parameters are estimated, and which sample is used. These decisions, therefore, affect comparisons of QI reliability-adjusted rates by hospital characteristics. Based on these findings, we recommend the following:

1. AHRQ should consider adding flexibility to the QI software to give users the option to use the reference population or reestimate the parameters for the risk- or reliability-adjustment models.
2. For users with large analytic samples that are substantially different from the reference population, we recommend strongly considering reestimating as many risk-adjustment and

shrinkage parameters as the sample size allows. Careful consideration is needed before reestimation to determine that the analytic samples differ in meaningful ways (for example, by risk in the patient populations or by hospital characteristics) that would necessitate reestimation and that within sample comparisons are the objective of the user (as opposed to using the current all-payer national reference population as a benchmark).

3. We recommend further analysis to determine if reestimating shrinkage parameters separately by hospital type or peer groups represents an improvement to the accuracy of estimated QI rates. Because shrinking to peer group targets can lead to substantial changes in QI rates by peer group, the approach warrants consideration. However, the findings in this analysis do not provide evidence regarding whether the differences by peer group reflect differences in quality or some other factor, which will largely determine whether the approach leads to improved accuracy of estimated QI rates. In addition, further study of alternate approaches to shrinking rates for small hospitals is warranted. Although the rates for small hospitals are most affected by the use of alternate shrinkage targets, because the confidence intervals are wide for small hospitals on average, it is the medium-sized and large hospitals that are more likely to experience a shift in performance categorization.

Target audiences

Reestimation of the QI software parameters and shrinkage to alternate targets could be particularly desirable to those conducting policy evaluations or implementing policy initiatives targeted to specific groups of hospitals or discharges using the QIs. Applications that may benefit from recalibration include public reporting of cross-sectional QI rates or performance, evaluations of quality improvement initiatives, pay-for-performance initiatives, or formal comparisons of different hospital groups.

Considerations for implementation

The feasibility of the modified approaches depends on the size of the analytic sample and the technical and programming capabilities to reestimate the shrinkage parameters and risk-adjustment models. First, these approaches are best suited for large datasets. For example, reestimating signal variance when the sample of hospitals is small or the number of discharges at each hospital is small can be problematic computationally. In addition, the ability to modify the software is a potential limiting factor for many users in implementing the modified approaches. Reestimating only the mean is relatively easy to implement; users estimate the mean of their sample and set it as the shrinkage target (prior mean) in the software. Reestimating signal variance and calculating the reliability weights for use in reliability adjustment is more complex. Reestimating the risk-adjustment models and integrating the results into the calculation of hospitals' rates is even more challenging and resource intensive. Thus, if the modifications to the approach are determined to be a valuable option for those who use the QIs, implementing the modifications may require AHRQ to add the necessary flexibility to the QI software to reestimate some or all of the parameters needed to estimate hospitals' rates (while maintaining the ability of users to run the software to calculate observed, risk-adjusted, and reliability-adjusted as currently specified).

The decision whether to reestimate the shrinkage parameters should consider empirical relationships in the analytic data. First, if the ratio of observed-to-expected rates is much different than 1, then the user should consider reestimating the risk-adjustment models alongside the shrinkage parameters. If the analytic sample permits and the methods are available, the risk-adjustment models should be reestimated on the analytic sample (Ash et al. 2011). An observed-to-expected ratio much different from 1 may indicate excellent overall performance, but it more likely represents poor risk-adjustment model fit or a sample that is much different from the reference population. Next, if shrinking to peer-group means, if one peer group contains a small set of providers, the researcher should consider whether the shrinkage target or signal variance is stable. Bootstrapping is one method to determine the stability of the shrinkage parameters. As a rule of thumb, reestimating only the shrinkage target for peer groups should be considered before moving on to reestimating the mean and the signal variance (Ash et al. 2011).

Remaining unknowns

The magnitude and direction of the effect of the shrinkage approach on the distributions of hospital rates, reliability weights, and performance categories are influenced by our analytic sample and do not necessarily represent the relationships in other samples. The results based on the select subsample are moderate because the subsample is large and fairly representative of the reference population. The external subsample of Medicare discharges also shows moderate effects, and we hypothesize that these effects would be larger for external samples that are smaller, cover a more homogenous group of patients or providers, represent a time period that does not overlap with the reference population, or contain data or QI specifications much different from the reference population. More work is needed to understand the stability of shrinkage parameters when sample sizes are moderate or small.

Future analyses

We recommend further analysis to determine if the effects of the modifications observed in the analysis represent improvement to the accuracy of QI rates and hospital comparisons. The simulation approach discussed in Chapter V provides the basic framework for conducting such an analysis. We also recommend further analysis on the potential of shrinking risk-adjusted rates using other information, such as a hospital's rate in a previous year or using information from other QIs for a given hospital to reliability adjust a rate for a different QI.

The QI composite indicators are risk- and reliability-adjusted, and because composites are often the focus of program evaluations and policy initiatives, more work is needed to understand the downstream effect of shrinking to alternative targets on compositing. Currently, the PSI, IQI, and PDI composite indicators are reliability-adjusted, and these adjustments are based on the same shrinkage parameters used to calculate reliability-adjusted rates.

This analysis modified the existing empirical Bayes model used to shrink the QI rates, but other modeling approaches are available. It is unclear how model structure and the choice of the prior distribution influences hospital comparisons. The next chapter in this report examines how other empirical Bayes or fully Bayesian models may influence results.

VII. EMPIRICAL BAYES AND BAYESIAN FRAMEWORKS

A. Methodological challenge

An essential goal of the AHRQ QIs is the reliable estimation of hospital quality. For purposes ranging from quality improvement to public reporting, statistical inferences about hospital quality are based on estimated rates of events that reflect inpatient care and patient safety. Importantly, appropriate use of the AHRQ QIs depends on a clear and explicit statement of a statistical framework and the model assumptions within that framework.

Although the current estimation approach is closely derived from an empirical Bayes model (Morris 1983), there are several small departures in the implementation of this framework and the documentation of the implementation that could lead to confusion regarding the appropriate approaches to making inferences using QI rates. Such confusion could lead to incorrect inferences and spurious comparisons of hospital quality. For example, although not explicitly stated, the current methodology assumes the true QI rates (across hospital types) come from a single normal distribution. However, the current technical documentation posits a gamma “posterior” distribution for the reliability-adjusted rates without justifying this specification through the appropriate prior distribution.[20] Given a normal likelihood, the normal prior distribution begets a normal posterior distribution, not a gamma distribution.

In addition, the statistical properties and validity of the resulting reliability-adjusted rates depend on the validity of the modeling assumptions. The current assumption of normality might not be the best reflection of true hospital rates, particularly given the skewed nature of observed hospital rates. Furthermore, the current assumption of normality is relatively restrictive, resulting in substantial shrinking of small hospitals’ QI rates to the reference population mean. This is a relatively conservative way to compare hospitals; for example, many small hospitals’ reliability-adjusted rates are shrunken close to the reference population mean.

B. Potential improvement

We recommend that AHRQ adopt a formal empirical Bayes or Bayesian framework to estimate hospital QI rates. In particular, a formal framework would provide rigorous inferences based on posterior distributions of true QI rates and formalize the role of prior assumptions and observed data in supporting posterior inference. AHRQ could also achieve increased clarity of the QI methodology through the formal specification of the approach. An additional key benefit of this approach is that an explicit statement of assumptions enables users to better assess for themselves the model’s suitability given their populations and observed hospital data.

We also consider alternative prior distributions that may improve the accuracy of estimated reliability-adjusted rates. We investigated prior distributions that could provide a better fit to the hospital discharge data than the current assumption of a normal distribution. For example, the EDA demonstrates that the AHRQ QI risk-adjusted rates are correlated with hospital characteristics, such as teaching hospital status and bed size (Jones et al. 2014a). Additionally, the rates exhibit skewed distributions with heavy tails, suggesting the existence of outliers.

[20] The technical documentation can be found here: http://www.qualityindicators.ahrq.gov/Downloads/Resources/Publications/2013/Empirical_Methods_r.pdf

Moreover, the appropriate distribution could vary by QI or QI module. For these reasons, a single normal distribution for all hospitals and QIs may not be an appropriate assumption.

Analytic approach

Under the empirical Bayes framework, closed-form expressions for the reliability-adjusted rates currently exist only for few special cases. The normal distribution is one of those cases, but it is difficult to consider alternative distributions without the aid of numerical methods. In our study, we use Markov chain Monte Carlo (MCMC) numerical methods to test alternative prior distributions, despite the fact that no simple formulas exist for them.

Using MCMC methods, we estimated reliability-adjusted rates based on alternative prior distributions while maintaining the current approach of assuming a normal likelihood for the risk-adjusted rates. We used discharges from our 12-state sample of SID for all models. To characterize inferences based on the different prior assumptions about the AHRQ QI rates (namely, how they are distributed), we calculated hospital ranks based on the reliability-adjusted rates and assessed changes in ranks between the current and alternate approaches. In Table VII.1, the 10 candidate assumptions about the reliability-adjusted rates are summarized. In particular, candidate models 5 through 10 used a prior distribution other than the normal distribution with common mean. The pieces of the methodology that were varied and assessed were the: approach to estimating the posterior distribution (current approach, analytic posterior, or MCMC); the assumed prior distribution (see Wang et al. [2014] for more detail on the rationale for choosing the candidate prior distributions); and the data used to estimate the prior mean to test the sensitivity or robustness of the results to the year used to define the prior (HCUP reference population or our 12-state sample from 2009 or 2010). Model 1 represents findings from the current AHRQ methodological approach; models 2 through 4 modify the approach to include a formal statistical framework with explicit assumptions, although the prior distribution is still assumed to be normal; models 5 through 10 further modify the approach to consider alternate assumptions regarding the prior distribution, including a normal distribution with peer-group means, gamma and beta distributions that appear to better fit the skewed observed distribution of hospital rates, and a mixture of normal distributions.

For each comparison, we examined the comparison plot to decide whether the modifications had an effect on hospital rankings and whether the effects were small or large. We also calculated the correlations for the rates and ranks for each comparison. The detailed analytic approach can be found in Wang et al. (2014).

Table VII.1. Candidate models of reliability-adjusted rates

Model	Model estimate	Prior distribution	Data for prior mean
1	From AHRQ v4.5	N/A	44 HCUP SID (2010)
2	Analytic posterior mean	Normal, common mean	12 HCUP SID (2010)
3	MCMC posterior mean	Normal, common mean	12 HCUP SID (2010)
4	MCMC posterior mean	Normal, common mean	12 HCUP SID (2009)
5	MCMC posterior mean	Normal, peer-group means	12 HCUP SID (2010)
6	MCMC posterior mean	Gamma	12 HCUP SID (2010)
7	MCMC posterior mean	Gamma	12 HCUP SID (2009)
8	MCMC posterior mean	Beta	12 HCUP SID (2010)
9	MCMC posterior mean	Beta	12 HCUP SID (2009)
10	MCMC posterior mean	Normal mixture	12 HCUP SID (2010)

Note: Models with the same prior assumptions will be grouped together in comparisons. Model 1 estimates hospital rates for hospital in the 12 SID in 2010 using model parameters in the current AHRQ QI software estimated on discharges in the HCUP 44-state reference population.

HCUP SID = Healthcare Cost and Utilization Project State Inpatient Databases; MCMC = Markov chain Monte Carlo.

N/A = not applicable. The current AHRQ QI methodology does not explicitly state a prior distribution, although it implicitly assumes a normal distribution.

C. Findings

Based on the candidate models described in Table VII.1, we chose 10 illustrative pairwise comparisons to assess the approach to estimating the posterior distribution, the assumed prior distribution, and the year of the data used to define the prior mean, which we list in Table VII.2. Through these comparisons, we aimed to assess the performance of the models on different data sources; closed-form expressions versus numerical methods; specification of peer-group means; and the form of the distribution—in particular, its skewness. For each specification of the prior distribution, we calculated the posterior means (that is, reliability-adjusted rates) of all hospitals for each QI examined in this analysis.

We summarize the findings of each pairwise comparison in Table VII.2. In particular, we have the following key findings. The detailed results of the analysis are available in Wang et al. (2014).

- In comparison A, we show that although only 12 states were selected for our analysis and a slightly different estimation method than the current AHRQ approach was used, the reliability-adjusted rates and ranks based on the 12 states are close to those based on 44 states for most of the QIs. We do find that some QIs were affected substantially by using only the 12 states data, for example, death rate in low-mortality diagnosis related groups (DRGs) (PSI 2) and postoperative respiratory failure rate (PDI 9). For these QIs, the results from 12 states may not be generalizable to the full reference population

- In comparison B, we verified that the posterior mean estimated by a numeric method (MCMC approach) is equal to the theoretical posterior mean. Therefore, comparisons C-J, which are comparisons of models estimated via MCMC, are appropriate for statistical inferences.
- Comparisons C-J show the one-to-one comparison of the posterior means under two different prior distributions. Overall, the posterior means are all affected by the change of the prior. The magnitude of the effect depends on the comparison and QI.
- Comparisons C-J also show that the small hospitals are more likely to be influenced by the prior distribution. This is because small hospitals have relatively diffuse likelihoods, since data tend to be sparse. As a result, the prior plays a relatively more important role.
- As indicated in Ash et al. (2011), a potential improvement of the current reliability adjusted rates is to add hospital characteristics to the prior without changing the normality of the distribution. Comparison C specifically addressed this issue in our analysis. In most cases, the comparison plots reveal that QIs are separated into the peer groups by clusters or bands of hospitals in the plots.
- Comparison D and E show that changing the prior from the normal distribution to a skewed distribution such as a beta or a gamma would have big effect on the rates on the right tail of the distribution (higher rates of adverse events).[21] However, the choice between skewed priors (such as beta vs. gamma) has little impact on the posterior means (comparison F).
- Using 2009 risk-adjusted rates instead of 2010 rates to estimate the prior parameters under an empirical Bayes framework tests the assumption that the distribution of the true QI rates is the same for both years. The results from comparisons G-I show that the effect on the posterior mean rates are very small, especially when comparing hospital rankings, providing evidence that the results are not sensitive to small differences in the year from which data come for this purpose. This finding suggests that reestimating shrinkage parameters when there is a small difference in the timing of the data used to estimate the prior versus the posterior QI rates might not lead to large changes in hospital results. However, this results likely does not hold over long periods of time or over periods of time when there were substantial changes to the rates of adverse events for a given QI.

[21] Many of the rates in the right tail of the distribution are for small hospitals and have low reliability weights on average. As shown in the analysis of shrinkage targets, the choice of shrinkage target (mean of the chosen prior distribution) has a large effect on rates for small hospitals but little effect on performance classification because the confidence intervals for the rates are relatively large. The choice of shrinkage target has a larger effect on classification of medium-sized and large hospitals. Thus, the finding that there is a large effect on hospitals in the right tail of the distribution likely does not translate into large changes in performance classification.

Table VII.2. Comparisons between candidate models of reliability-adjusted rates and summary findings

Comparison	Models compared	Motivation for comparison	Key findings
A	1 and 2	Compare the use of different reference populations (current AHRQ 44 states and the 12-state study sample)	Small effect on most QIs and substantial effect on a few QIs
B	2 and 3	Compare the numeric posterior mean to analytic posterior mean	Little to no effect on the posterior mean rates
C	3 and 5	Explore the effect of peer group means in the prior distribution	Substantial effect on rates and rankings for most QIs; creates clusters of results based on peer groups
D	3 and 6	Explore the effect of a skewed prior distribution	Skewed posterior mean rates
E	3 and 8	Explore the effect of a skewed prior distribution	Skewed posterior mean rates
F	6 and 8	Explore the effect of using different skewed distribution	No substantial effect on posterior means
G	3 and 4	Explore the effect of year in prior estimation	Small effect on the posterior mean rates
H	6 and 7	Explore the effect of year in prior estimation	Small effect on the posterior mean rates
I	8 and 9	Explore the effect of year in prior estimation	Small effect on the posterior mean rates
J	3 and 10	Explore the effect of a normal mixture prior	Substantial effect on posterior mean rates

Source: Mathematica Policy Research analysis of SID data from 12 states from January 2009 to December 2010 using AHRQ software version 4.4.

Note: The substantial, small, and no effect are based on visual examination of the comparison plots

D. Recommendations

Regardless of whether AHRQ adopts a formal empirical Bayes or Bayesian framework in place of the current approach, clarity of exposition with explicit assumptions stated is crucial to ensuring that users are able to correctly estimate QI rates and draw inferences from the findings. Regarding the current assumption of normally distributed hospital rates, modifying this assumption leads to substantial changes in the rates. Currently, there is not enough evidence to state if the changes represent improvements or which alternate distribution should be selected for various users of the QI software. However, the observed distributions of rates for some QIs suggest that the current assumption of a normal distribution should be reconsidered. We recommend model checking exercises in this section to assess the fit of models to the hospitals' data and the distributional assumptions of the prior to help choose the most appropriate prior distributions.

Target audiences

Users of the AHRQ QIs will benefit from statistical inferences that are based on clearly and explicitly stated assumptions about the prior distribution of the AHRQ QI rates. In particular, the Bayesian framework, facilitated by MCMC estimation methods, sets the stage for further study of the AHRQ QI methodology by providing a flexible statistical framework in which assumptions can be stated and tested.

Considerations for implementation

If non-normal priors are adopted under the Bayesian framework, AHRQ will need to build MCMC routines into the software so that it can produce results based on simulating a posterior distribution. To implement the Bayesian model, future development of the software will involve testing candidate modules with alternative specifications of the prior distribution. An important component of development, particularly for user-acceptability testing, should involve mixing and convergence diagnostics, which will enable AHRQ to ensure that the MCMC numerical methods are producing valid estimates. In particular, the length of the MCMC chains requires careful specification, with longer chains providing more accurate estimation of posterior means.

Based on the simulated draws from the posterior distribution, AHRQ could perform elementary calculations to estimate the posterior means and variances. Ultimately, the full posterior distribution for each hospital could be reported as well. User guides would need to clearly specify how to use the full posterior distribution for making statistical inferences.

Remaining unknowns and further analysis

As with any Bayesian analysis, the prior distribution encapsulates our beliefs and assumptions about the behavior of the true, underlying hospital QI rates, which are essentially unknown. Due to the importance of the prior distributions, we recommend that AHRQ consider incorporating additional data resources, such as HCUP data over several years, to determine the specification of the prior distribution, such as its skewness or multimodality. We also recommend further analysis of semiparametric or nonparametric prior distributions, which would enable the data to govern the shape of the distribution, instead of explicit assumptions set by AHRQ or the end user.

A common critique of reliability adjustment in any hospital quality measurement model is that shrinking to a single mean likely masks the underlying variation in quality, so that effectively, any hospital with a small number of discharges cannot differ from the average. One of the approaches that could avoid over-shrinkage of the risk-adjusted rates is to consider a multivariate limited translation hierarchical Bayes estimator (Ghosh 2011). This estimator considers hospital size and the distance between the hospital mean and the national mean in determining the magnitude of the shrinkage. Another potential approach is to consider using additional data resources to gather more information about the prior for small hospitals to address the concerns and provide solid inferences. This method could use data from prior years to inform the shrinkage targets for small hospitals.

An important step in any statistical modeling activity is to assess the fit of the model to the data and to the analyst's substantive understanding of the analytic goals. This assessment is especially important for Bayesian analyses, for which it is imperative to ensure that the likelihood reflects the underlying data and that posterior inference is not being driven by prior assumptions. Based on this statistical principle, we recommend that AHRQ apply formal model checking techniques to any models under consideration, including the current AHRQ QI models. Formal model checking would possibly mitigate concerns about the sensitivity of reliability-adjusted rates to prior assumptions, including the assumption of normally distributed risk-adjusted rates. For consideration, we suggest two main statistical approaches to model checking: through posterior predictive checks and cross validation (see Wang et al. 2014 for more detail regarding these approaches to model checking).

As a proof of concept of such a model-checking exercise as those mentioned above, we simulated replicated data under the empirical Bayes analogue of the current AHRQ model (Model 3 from Table VII.1). These replicated datasets describe the data that might have been collected if the model were true. The most fundamental way to check model fit is to compare such replicated datasets to the actual data (Gelman and Hill 2007), with systematic differences between data and replications indicating poor model fit. As an illustrative example of the model deficiencies that can be discovered using this type of “posterior predictive” check, we note that in reality there are many hospitals with risk-adjusted QI rates of zero and that of course there are no hospitals with negative rates. In the replicated datasets of risk-adjusted QI rates, by contrast, we see no zeroes and many negative values; there are also few values very close to zero. These findings demonstrate that not only should AHRQ consider alternate prior assumptions for the true hospital QI rates as we have discussed above, but that the assumption of a normal likelihood may merit careful evaluation as well. It also confirms our recommendation that AHRQ adopt a rigorous statistical framework to allow careful model checking and improvement.

VIII. DISCUSSION

The analyses conducted under this project provide evidence supporting a range of recommended steps for AHRQ to improve the accuracy of hospital comparisons made using QI rates. Table VII.1 summarizes the recommendations in the four areas of analysis: the standardization approach, incorporating hospital characteristics in the risk-adjustment models, estimating alternate shrinkage targets, and implementing an empirical Bayes framework.

We found small effects on most differences by hospital type in mean QIs when the method of risk adjustment is changed from indirect to direct standardization. However, for several QIs, primarily PDIs, the conclusions drawn regarding differences in mean rates by hospital type changed under direct standardization. In addition, five percent or more of discharges could not be matched across hospital types for several QIs, which could influence the findings. For such indicators, direct standardization using a common risk profile across hospital types may be preferred. Though the effect of hospital type on such QIs can be substantial and accounting for it may change the results of comparisons, our simulations suggest that incorporating this information in risk adjustment itself is not desirable. Similarly, classifications of hospitals' performance using reliability-adjusted QIs are meaningfully affected by the decision to estimate parameters within peer groups defined by hospital type or a particular study sample and by the choice of the form of the prior distribution. For that reason, we recommend that the flexibility to estimate such parameters and choose alternate prior distributions be added to the QI software. However, we cannot conclude that the results using these reestimated parameters and alternate prior distributions are more accurate. Although, the framework for evaluating the appropriateness of alternate models can be established using an explicit empirical Bayes or Bayesian framework to estimate the models and QI rates.

Our strongest recommendation, which comes from the empirical Bayes analysis, is to formalize the statistical framework for estimating QI rates with clear, explicit assumptions (such as the form of the prior distributions and likelihood) tied to the framework. The formalization of the framework will add support to inferences drawn from QI results (such as hospital comparisons) by producing QI rates from explicit and testable distributional assumptions. A concise framework with clear documentation provided by AHRQ will also add to the clarity of the overall approach for all users.

Table VIII.1. Summary of recommendations to improve the methodological approach and remaining unknowns

Area of analysis	Recommendations	Remaining unknowns
Risk adjustment		
Direct standardization	Consider direct standardization for user aiming to compare performance on a specific set of patients.	Differences in directly standardized rates by hospital type could be driven by differences in unmeasured risk or differences in quality, just as indirectly standardized rates (that is, direct standardization does not overcome the concern of omitted risk factors).
Incorporate hospital characteristics in risk-adjustment models	Do not incorporate hospital type indicators in the risk-adjustment models as risk factors. Further analysis of including discharge-level variables in the risk-adjustment models as potential proxies for unmeasured risk is needed. If there is evidence of unmeasured risk and the appropriate discharge-level variables are not available, consider stratifying results by hospital characteristics.	The extent to which various factors contribute to the differences in QI rates by hospital type.
Reliability adjustment		
Shrinking to alternate targets	Consider adding flexibility to enable users to reestimate the shrinkage parameters and risk-adjustment models.	It is not clear whether changes in hospital rates due to shrinking to peer group means reflect improvements in the accuracy of QI rates.
Incorporating a formal empirical Bayes framework	Implement a formal statistical framework with explicit assumptions and clear documentation. Consider alternate distributional assumptions for the prior and likelihood (other than the current assumption of a normal distribution).	The choice of distribution will depend on the empirical evidence regarding observed hospital distributions. The choice could depend on the user's sample of discharges and hospitals.

In addition, the findings in the empirical Bayes analysis support additional tests to determine whether AHRQ should invest in allowing the flexibility of distributional assumptions other than a normal distribution. The normal distribution may be too restrictive in light of evidence of skewed distributions for some QIs and given that the normal prior leads to a substantial degree of shrinkage of hospitals at the extremes of the observed distribution to the mean of the reference population. If the normal assumption substantially mischaracterizes the distribution of hospital effects, it will lead to incorrect measurement of quality and incorrect inferences about hospital performance.

It is more challenging to make a definitive recommendation regarding the most appropriate distribution to use in the QI models. We recommend further analysis to examine the distributional assumptions that are appropriate for various patient and hospital populations. Based on the findings from these analyses, AHRQ could consider adding the flexibility to use alternate distributions allowing the data to guide the choice. We also recommend further analysis on the distribution that best fits the national reference population and the extent to which implementing this distribution improves the accuracy of QI rates over the current use of the normal distribution.

It is particularly challenging to make definitive recommendations regarding modifications that incorporate hospital characteristics in AHRQ's methodological approach. For example, the analysis of incorporating hospital type indicators as risk factors in the risk-adjustment models provides evidence that the modification leads to an improvement in very limited circumstances (that is, when nearly all of the differences in hospital QI rates can be attributed to unmeasured risk). However, because we cannot be confident regarding the extent to which unmeasured risk contributes to differences in hospital QI rates, we cannot say definitely that AHRQ should not consider this modification under any circumstances. Further analysis is needed to better understand the role that unmeasured risk plays in the observed differences in rates across hospital types and the point at which incorporating hospital type indicators represent an improvement to the accuracy of estimated rates to make a definitive statement. Until this role is better understood, we recommend stratification of results by hospital type when the concern that indicators harm hospitals in a particular peer group is paramount; although, we recommend considering results within hospital type jointly with results across hospital types to maintain the comparison of performance to a national benchmark. In addition, in light of the current findings and the relative promise of incorporating discharge-level proxies for unmeasured risk as an alternative approach to accounting for differences in unmeasured risk by hospital type, we recommend that AHRQ consider additional research on inclusion of hospital-level variables as risk factors a low priority. We discuss potential future analyses of incorporating additional discharge-level factors later in this chapter.

The uncertainty surrounding factors that contribute to differences in hospital QI rates also makes it difficult to make a definitive recommendation regarding the suitability of shrinking rates using peer group shrinkage parameters. Our analyses demonstrate that there are differences in mean hospital rates by hospital type, and that even when the magnitude of these differences is large, shrinking to different means for the hospital type leads to small changes in hospital classification by the Hospital Compare approach. However, further analysis is needed to determine if these changes represent an improvement in the accuracy of hospital classification. If the differences in mean rates reflect differences in quality by hospital type, shrinking to peer group means is appropriate. If the differences in mean rates reflect differences in unmeasured risk across hospital types, shrinking to the peer group means will likely exacerbate the issue by further shrinking hospitals of each type to rates that are different due to unmeasured risk, not quality. Simulation testing analogous to that used to test changes to risk adjustment would help us to assess the circumstances under which shrinking to peer group means represents an improvement.

The reestimation of shrinkage parameters within sample when the analytic sample is substantially different from the HCUP reference population shows promise in improving the accuracy of QI comparisons. We recommend that AHRQ consider adding flexibility to the QI software to allow users to reestimate the risk- and reliability-adjustment models to better fit the sample of hospitals they are analyzing. This flexibility could enable users to incorporate their own data to produce parameters that better fit the patients relevant to them. For example, shrinking hospital rates for care of Medicare fee-for-service patients to the mean rate for all Medicare fee-for-service patients could lead to rates that show variations in treatment of that patient group more accurately than shrinking these rates to the all payer and all age HCUP reference population. Furthermore, reestimating risk-adjustment models within sample creates risk-adjusted rates that compare a hospital's performance on their patients to the expected performance of an average hospital within their population of hospitals rather than an average hospital in the HCUP reference population; which could be a more appropriate comparison if the set of hospitals and patients is substantially different from those in the reference population. However, users should carefully consider the conceptual rationale and empirical evidence supporting reestimation for their sample before making the decision. Further analysis is needed to determine if reestimating the parameters on various analytic samples represents an improvement in the accuracy of estimated rates. It is also worth reiterating that this approach is only computationally feasible when the analytic sample contains a large number of discharges in order to reestimate the models and parameters.

We also compared QI rates by hospital type across the same average patient population using direct standardization and found that the conclusions were the same as those drawn from rates using the current method of indirect standardization, with a few exceptions. If direct standardization produces risk-adjusted rates with tighter control of the observed risk factors, the rates could provide better estimates of the differences between two hospital types. In this case, the few instances in which the differences estimated by hospital type vary according to indirectly and directly standardized rates should be examined further to determine what might be driving the differences and whether the finding is evidence that a change is warranted to the risk-adjustment models for these QIs. Also, because we found that results for some QIs differ meaningfully when the standardization method is changed due to differences between the risk profiles of different hospital types, direct standardization to the risk profile of one hospital type may be desirable. However, we should be careful to note that directly standardized rates suffer from some of the same key potential drawbacks as indirectly standardized rates; namely that there could be unobserved risk factors that differ by hospital type and contribute to the observed differences in rates by hospital type. Further analysis is needed to understand and account for any differences in unmeasured risk whether the approach is indirect or direct standardization. In addition, although not directly assessed in the analysis, we recommend that AHRQ further examine the suitability of direct standardization as a potential method for calculating risk-adjusted rates for the QIs (which could utilize the same clinical specifications as the current approach) for users that aim to compare hospitals' performance on a specific set of patients.

A. Overarching recommendations when deciding how to approach hospital characteristics

Given the difficulty in determining the extent to which specific factors such as unmeasured risk contributed to differences in hospital rates by hospital type, we recommend that any decisions regarding the inclusion of hospital characteristics and how to define them begin with a strong conceptual rationale. That is, it should be determined whether there is a strong clinical foundation for why patients that visit a type of hospital face a greater risk of an adverse event that is not related to the quality of care delivered at the hospital type before ever considering adding a hospital characteristic in the estimation of hospital rates.

In addition, although there were no specific end uses or users targeted when conducting these analyses, we consider the suitability (drawbacks and advantages) of different approaches and modifications for several broadly-defined users of the QIs with the following objectives: (1) hospital quality improvement, (2) direct hospital comparisons across hospital types, and (3) patient decision-making regarding site of care. We summarize the issues associated with each approach for each end user in Table VII.2. For example, the current approach using indirectly standardized rates, which compare a hospital's performance for their patients to the expected performance of an average hospital, is likely appropriate for a hospital targeting areas for quality improvement. An ability to combine this information with unadjusted rates by patient type could also be beneficial for this objective. Patients comparing local hospitals to decide where to have a procedure performed will require different information and comparisons; perhaps comparing all candidate hospitals' expected performance for patients similar to them using a direct standardization approach (although because many of the events captured by the QIs are rare in such hospitals, the number of discharges may be insufficient to support this approach). Again, patients could also benefit from seeing unadjusted rates for different patient types. Although, providing unadjusted rates over a longer time period might be necessary for small hospitals with relatively unreliable rates to avoid misrepresentation of performance estimated over a small number of discharges; thus, it is important to consider the unadjusted rates jointly with their overall reliability-adjusted rates. Lastly, an approach that attempts to adjust hospital rates for unmeasured risk is likely more desirable if there is a concern that a program using the QI results will unjustly penalize hospitals with riskier patients, such as a federal program to adjust payment based on QI rates. The lack of an adjustment could lead to undesirable changes in admitting practices based on risk and targeting of low-risk patients, or could undermine the financial status of hospitals serving risky populations. However, addressing this concern through adjusting hospital rates by type would also have the "side effect" of adjusting away quality by hospital type, thereby nullifying the incentive to expend efforts to improve quality for a certain type of hospital, such as increased dedication of resources to reduce patient safety events at hospitals serving high proportions of low-income patients.

Although the QIs are not designed and maintained for any one specific user or even specific use, AHRQ can provide added flexibility for users to better align the approach and the end uses. AHRQ can contribute to these decisions by adding flexibility to the methodological approach to allow for the incorporation of hospital characteristics in several ways: for example, estimation of reliability-adjustment parameters by peer group or for the analytic sample or direct standardization within user-defined peer groups. AHRQ can provide the guidance and tools to facilitate these variations on the current approach (that is, the specifications needed to define eligible cases and adverse events, model specifications to allow reestimation, and clear methods documentation).

Considering the uncertainty surrounding the factors behind differences in QI rates by hospital type, difficult policy decisions will need to be made by the users of the QIs and other quality measures. The decision regarding whether and how to consider hospital characteristics is a policy decision that should be made after carefully considering the objective of the program's use of the QIs, conceptual rationales for the relationships between hospital characteristics and the QIs, and all available empirical evidence regarding the relationships and the appropriateness of various approaches.

Table VIII.2. Methodological considerations for users of the QIs

Objectives of quality measurement	Methodological considerations across four areas of study
Hospital quality improvement	• **Standardization approach**. Indirect standardization provides comparisons of a hospital’s performance with an average hospital on the hospital’s population, which is appropriate for this purpose. • **Risk-adjustment models**. In addition to comparison of risk-adjusted rates to a national benchmark across all hospital types, it could be helpful for hospital’s to compare their rate within their hospital type to identify areas for improvement. However, for comparisons within type we recommend a stratification approach rather than incorporation of hospital type indicator in the risk-adjustment models. In addition, because a hospital focusing on internal quality improvement might not be as concerned with how their differential patient risk influences their rate of adverse events (that is, they want to improve care regardless of patient risk), it is advisable to jointly consider the adjusted and unadjusted rates when making decisions regarding how to allocate resources for quality improvement. • **Shrinkage targets**. Reestimating the shrinkage parameters within sample will not be feasible for a single or even a small group of hospitals. However, shrinking to a peer group target could be appropriate if quality of care differs by hospital peer group. Because hospitals’ reliability-adjusted rates could be quite different than their unadjusted rates (particularly when hospitals have relatively few qualifying discharges), it is advisable to consider rates with and without reliability adjustment.
Large-scale comparisons across hospital types	• **Standardization approach**. Indirect standardization provides comparisons of all hospital types to a hypothetical average hospital. This is an ideal approach if there is enough overlap in the hospital populations of interest with the HCUP reference population. If not, a form of direct standardization could be preferable. • **Risk-adjustment models**. To the extent possible, adjustments should be made to account for all differences in patient risk factors to increase the equity of comparisons. Given the uncertainty surrounding the extent of unmeasured risk, an adjustment by type might be more attractive for a user that is concerned about unfairly punishing hospitals with higher proportions of high-risk patients under the current approach, but the tradeoff is obscuring differences due to different quality between types. • **Shrinkage targets**. If there are differences in quality by hospital type, peer group means could represent more appropriate shrinkage targets for reliability adjustment. • **Empirical Bayes framework**. A formal statistical framework with the flexibility to allow distributional assumptions in the model other than the normal distribution is to the benefit of all users of the QIs interested in comparing risk- and/or reliability-adjusted rates across hospitals. The choice of the distributional assumptions will depend on the underlying true distribution of the hospital rates for the population of hospitals and the QIs to be examined.
Patient choices regarding site of care	• **Standardization approach**. Patients considering their site of care could benefit from comparing multiple local hospitals to a single average hospital based on the patients actually treated by these hospitals. However, direct standardization could provide a comparison of the hospitals for the same patient population or hospital type, which could be a population that is of specific interest to the patient (for example, high risk patients with multiple comorbidities). Although sufficient sample size for subsets of patients would likely be an issue for producing this for small hospitals. • **Risk-adjustment models**. To the extent possible, adjustments should be made to account for all differences in patient risk factors to increase the accuracy of information available to patients. Otherwise, they could select a hospital unknowingly because it has a lower proportion of high-risk patients. Patients could also benefit from unadjusted rates for patients similar to them. • **Shrinkage targets**. Shrinkage of unreliable rates could help prevent patients being misled by unusually low or high rates that are largely due to random variation rather than quality. If there are differences in quality by hospital type, shrinking to peer group means could improve the accuracy of rates available to patients. Patients could benefit from considering these reliability-adjusted rates in combination with the unadjusted rates or observed rates over time, as the reliability-adjusted rates could also obscure the quality of hospitals that truly have performance in the tails of the distributions.

B. Comparisons of hospitals within hospital type

One of the objectives of our research is to identify situations in which differences in QI results by hospital type indicate that comparisons of QIs across hospital type should not be made. We have identified three scenarios: comparisons are not valid because differences in QIs do not indicate differences in quality when hospitals of different types are being compared; comparisons are valid but harmful because hospitals of one type are performing a role that cannot be performed otherwise; and comparisons are not valid because hospitals are not providing the same treatment.

Our research has provided some evidence against each of these arguments. In particular, our simulation results suggest that, even when differences by hospital type are associated primarily with patient risk, ranking hospitals of different types together provides valid quality information. Risk adjustment controlling for hospital type is misleading because a hospital's performance relative to its type may be compared directly with another hospital's performance relative to a different type. If we are concerned about the validity of differences between hospitals related to their characteristics, the evidence suggests that we should look for patient-or discharge-level variables that can account for them. If we are concerned about potential harmful effects of comparisons across hospital types, either because of the risk of mistaking risk differences for quality differences, or the risk of harming hospitals serving a disadvantaged population, providing education or testing new technologies, stratification rather than risk adjustment by type is preferred because it makes explicit that cross-type comparisons are prevented. However, even when direct comparisons by hospital type appear to be harmful, other means than stratification, such as compensating subsidies or special assistance to hospitals according to their social function, may be preferable.

The third scenario, in which comparisons should not be made because different hospital types offer different treatments, is one for which our comparisons of direct and indirect standardization provide some evidence. Direct standardization differs from indirect because it forces the comparison between hospitals of patients that are similar in known dimensions rather than comparing the outcomes of one hospital treating its own patient mix to other hospitals treating their own patient mixes. If, when we compare directly standardized differences by hospital type with indirectly standardized, we find that many groups of patients cannot be directly compared because they are only treated at one hospital type, or if performance by hospital type looks very different when measured by direct and indirect standardization, it would suggest that specialization of hospitals in treating certain categories of patients makes comparisons across hospital type invalid. When we compared the results of direct and indirect standardization, we identified some cases in which the scenario of imperfect comparability may hold. For most IQIs and PSIs and the hospital characteristics we reviewed, we encountered neither problem. However, for IQI 11, PSI 12 and several PDIs, we found notable differences in patient populations and/or comparisons of QI rates by hospital type. Together, these findings suggest that stratification of some QIs using the hospital types we identified or using some other method may be advisable, because there may be significant differences in the treatment they provide.

These issues are complicated by the use of shrunken or reliability-adjusted rates. The purpose of shrinkage is to minimize the error in the estimation of rates given the information at our disposal. However, the estimates are weighted averages of the hospital's own rate with the peer group rate. The implications of the choice of whether to include the hospital's own characteristics in the shrinkage process are in large measure opposite to the implications when choosing to include hospital characteristics in risk adjustment. Shrinking to a single national average amounts to the prior assumption that all hospital types have the same mean quality. Therefore, if one is unwilling to attribute differences by hospital type to unmeasured risk, one might prefer to include hospital characteristics in shrinkage. On the other hand, the assumption that differences are associated with quality differences, as implied by setting up different shrinkage targets, is equally undesirable in the absence of definitive evidence that hospital types differ in mean quality. If shrinkage is used, stratification and the avoidance of comparisons outside of a peer group in instances in which hospital comparisons would be altered by using a different choice of shrinkage target may be the desired solution.

Though the results of comparisons by hospital type do not suggest that these types delineate peer groups performing different clinical roles for most QIs, such peer groups may still exist. In addition, although large and teaching hospitals are generally those with the most complex patients, some hospitals of other types also treat these patients. Therefore, peer groups might cut across those categories. Additional research could identify peer groupings based on clustering of patient characteristics, severity, or hospital characteristics, and separate models could be estimated within peer group. This form of peer grouping is often used in efficiency profiling and is consistent with our investigation of shrinkage to alternative targets, in which we conclude that reestimating parameters using the peer group defined by a study sample, or by peer groups within sample, may be appropriate.

Other than the flexibility afforded by the ability to estimate parameters from within sample already discussed above, these findings do not suggest any changes to the QI software. Stratification can be performed simply by separating the results or the input data into peer groups. However, if cross-cutting peer groups are identified, more extensive changes to the indicators would be entailed. Such peer groups might be used to restrict the denominators of indicators to certain types of patients or create separate QIs applicable to specific groups of hospitals.

C. Extensions of current analyses and recommended next steps

The findings generated from the analyses suggest additional topics for investigation. These additional studies would focus on factors that might explain the relationships between QIs and hospitals characteristics and on methodological improvements to the QIs. First, we recommend logical extensions of the current analyses. We expand the scope from that of this project to consider modifications to discharge-level variables in risk-adjustment models and more generally about improvements in the accuracy of rates (that is, not necessarily just for the purposes of improving hospital comparisons, although any improvements in the accuracy of rates should also improve the accuracy of hospital comparisons and hospital comparisons across hospital types). We also expand the scope to consider methods other than risk and reliability adjustment. In particular, we also recommend analyses to examine potential improvements to: (1) methods used to estimate the AHRQ composite indicators, (2) the joint estimation of risk- and reliability-adjusted rates, and (3) approaches to using QI rates to profile hospital quality.

Extensions of analyses on risk- and reliability adjustment

Examining factors contributing to differences in QI rates by hospital type. As discussed throughout the report, it is difficult to uncover evidence regarding the extent to which different factors, such as patient characteristics, processes of care, and structural quality contribute to the differences in QI rates by hospital type. Recommendations to improve the AHRQ QIs can be made more confidently and specifically as evidence regarding these factors is uncovered. Modifications can then be designed based on a specific and well understood threat to validity of the rates rather than hypothesized threats. Any analysis that contributes to this understanding should be made a priority for a research agenda regarding risk-adjustment methods. We briefly summarize analyses that could improve our understanding of the factors driving the differences in QI rates by hospital type and better inform decisions regarding the potential modifications in Appendix A. The analyses range from sensitivity analyses that inform the likelihood that an unmeasured risk factor could explain the observed differences to enhanced matched case-control and instrumental variable approaches that attempt to isolate differences in quality from other factors.

Standardization approach. Several approaches to direct standardization hold promise for improving the estimation of QI rates for users that aim to compare hospitals for the same set of patients. Template matching, described in Silber et al. (2014), is one such approach that warrants further study for applications of the QI rates. In this particular approach to template matching, each hospital is compared to other hospitals based on their patients; in effect, the approach is a mixture of indirect and direct standardization. In another application of template matching for comparisons within hospital types, hospitals could be compared based on a hospital type template; for example, teaching hospital would be standardized based on a teaching hospital template of patients and nonteaching hospitals would use a nonteaching template of patients. The current analysis on standardization can be modified and extended to consider these various approaches to direct standardization.

Inclusion of discharge-level factors correlated with risk. We identified consistent differences in QI results according to factors that are indicative of the SES of hospitals' patients, such as DSH status or Medicaid enrollment. Our results also suggest that incorporating patient- or discharge-level characteristics in the risk-adjustment models is a more promising approach than including hospital type indicators. Those arguing for inclusion of patient-level SES (for similar reasons as those posited in this report for including hospital characteristics) have focused largely on readmission measures, not the AHRQ QIs. The argument is that some patients are at a greater risk for readmission due to their health-related behaviors outside of the hospital, hospitals cannot be held accountable for these behaviors, and hospitals with disproportionate shares of these patients will be unjustly penalized. In contrast, the AHRQ QIs seek to measure events that happen during inpatient stays. The conceptual rationale for the link between unmeasured risk and mortality could be well-founded if healthy behaviors outside of the hospital cannot be observed or are not properly accounted for by other demographic and health characteristics already included in the models and these behaviors influence the likelihood of mortality in the hospital. It is more difficult to argue that certain patients are at greater risk for patient safety events occurring in the hospital after accounting for the current set of risk factors, but that differences do not reflect the quality of care they receive. That is, patients with similar demographic and health characteristics receiving the same quality of care should face the same risk of a patient safety event. The exception to this claim occurs if there is some aspect of their health and risk that is not accounted for by the current set of risk factors; for example, low SES could be correlated with factors such as increased stress, higher rates of obesity, or environmental pathogens, that could lead to elevated risk of certain patient safety events (the SID include flags for obesity, and the flag is included as a risk factor for a subset of QIs). We recommend further analysis to identify any conceptual rationales for differences in risk by SES and attempt to validate the rationale empirically. In addition, factors other than SES should be considered to account for patient risk, such as whether there are clinical or behavioral characteristics available that more directly account for patient risk, such as smoking, alcohol or substance abuse, or frailty.

A logical extension of the risk-adjustment analysis presented in Chapter IV is to examine the effects of incorporating *patient* characteristics contained in discharge data related to hospital characteristics and associated with outcomes in the QI risk-adjustment models on the accuracy of hospital comparisons. A critique of the general approach to risk adjustment for quality indicators has focused on the omission of certain patient characteristics correlated with risk in the discharge-level models. In particular, the argument has been made that patient risk varies by the SES of patients (even after accounting for the demographic and health characteristics of patients in the current QI models) and the proportion of patients with low SES varies by hospital and hospital type. Several of the hospital type effects identified in this report may be attributed in part to patient SES. Thus, the argument concludes that patient SES should be included as a risk factor in the models to account for unmeasured risk and improve the accuracy of the estimated QI rates; that is, the same argument highlighted throughout this report.

The challenges in a discharge-level version of the analysis are largely the same as the challenges in the current analysis. The primary challenge is determining whether the observed differences in the rate of adverse events for discharges with a characteristic are due to the risk generated by the characteristic or differences in care delivered to patients with the characteristic. In the SES example discussed above, it could be that low-SES patients have an elevated risk regardless of care, or it could be that low-SES patients receive lower quality of care on average (or a mix of both). An additional challenge to a discharge-level analysis is the availability of variables that indicate the SES of patients. In the SID, candidate variables include indicators for Medicaid as the primary payer and uncompensated care. Another option, in the absence of discharge-level SES, would be to include summary metrics indicating SES of the areas in which the patients live. Because the SID do not contain location information for patients besides state, an alternate data source would be required to facilitate this approach, such as Medicare claims data, which include patient zip codes.

Given the challenges outlined above, the first objective of the analysis should be to establish whether there is a strong rationale for including SES or similar variables in risk-adjustment for the various QIs and QI modules. To the extent possible, the conceptual rationale should be supplemented by empirical evidence illuminating the contributions of factors driving the differences (risk or quality). For example, the analysis could split the patients in each hospital by SES and compare risk-adjusted estimates of performance (a discharge-level matched control analysis could achieve a similar objective). The differences could still be influenced by differential treatment of low-SES patients within hospitals and the proportion of low-SES patients could also affect overall performance of all patients at the hospital, but this analysis will provide evidence of the factors contributing to the differences. The remainder of the analysis would proceed in a similar fashion to the current analysis: an examination of how the modification affects model performance and hospital-level results and a simulation analysis of how the modification performs under different explanations of the differences in rates. Ultimately, it may not be possible to verify an SES effect. In that case, quality indicators should not be distorted by including unvalidated risk factors. Rather, the design of programs should consider the likely effect of using the indicators across populations with a range of SES, but without compromising the incentives to improve quality contained in the QIs. Combining the information in stratified and unstratified rates across all patient and hospital types in the design of programs is one such possible approach.

Other potential improvements to the risk-adjustment methods. We also recommend additional analyses assessing potential improvements to the methodological approach to modeling risk, including: hospital-specific random effects in the risk-adjustment model to capture correlation across patients at the same hospital, and the complexity of the covariate structure in the models to consider interaction terms and the use of approaches such as classification and regression trees and random forests to improve the case mix adjustment (Ash et al. 2011).

Exploring potential improvements to shrinkage approach. Hospitals with limited numbers of qualifying discharges for the QIs do not provide enough information to generate rates that confidently represent quality rather than chance, and so present a considerable challenge to estimating rates. We recommend further study of the potential improvements to reliability-adjusted rates from borrowing strength from alternate sources, such as: (1) shrinking hospitals' rates to their rates in previous years and (2) shrinking rates for one QI based on the information provided in other similar QI rates for the same hospital. We also recommend further analysis of approaches to limiting the amount of shrinkage, even for the smallest hospitals, such as limited translation hierarchical Bayes estimators. Furthermore, it might also be prudent to consider a wholly different approach to measuring performance on particularly rare QIs below a certain number of qualifying cases, such as a the number of days since the last event.

Recommended analysis of other components of the QI methods

In addition to extensions of the analyses on the AHRQ risk- and reliability-adjustment methods, we considered the implications of modifications to the approach to estimating values for the composite indicators, the overall approach to risk- and reliability-adjustment in estimating QI rates, and the ways that the QIs are used to compare hospital quality.

Exploring methodological improvements for the composite indicators. Many of the current uses of the AHRQ QIs for hospital comparisons make use of the PSI 90 composite indicator (safety for selected indicators). The EDA established that the composite indicators have similar relationships with hospital characteristics to their component indicators. Because of the importance of the composite indicators in hospital comparisons and the demonstrated relationships with hospital characteristics, a logical extension of our QI methods research is to examine the effect on composites of changes in methods. One option is to begin with the modifications to component indicators we have tested and assess the "downstream" effects of modifications on the composites they comprise. The current compositing methodology combines the reliability weights built into the reliability-adjusted component estimates with prevalence weights, such as numerators or denominators. Thus, if components are shrunken to alternative priors arising from different hospital characteristics, the composite impact of the change would depend on the extent to which these priors embodied a consistent positive or negative relationship between the component indicators and that characteristic. Another change with potential downstream impact is estimating signal variance by peer group, which would change the distribution of reliability weights and consequently the composition of the composite. Like their component indicators, testing results also may suggest that composites are best used for comparisons within peer group.

In addition, future work could address modifications to the methods used to construct the composites, in particular, approaches to weighting. Alternatives to prevalence weights include weights based on the salience or positive predictive value of the QIs, which may vary according to hospital type. Another approach that might be tested is to use a Bayesian framework to reliability adjust the composites directly rather than adjusting the component indicators as is currently done. A related extension of the analyses in this report is to investigate the effects of hospital characteristics on quality indicators by taking advantage of variation over time and across QIs to identify latent quality or risk effects. The results can serve as evidence for hospital type effects or as the basis of revised composite weights and groupings.

Exploring a unified risk- and reliability adjustment approach. Relative to AHRQ, CMS adopts a similar but modified approach to obtain reliability-adjusted estimates of hospital quality. Instead of using a two-stage model in which the risk- and reliability-adjustment steps are separate, a unified hierarchical logistic model is used to fit the discharge-level data. This unified approach shrinks model parameters as part of a single estimation procedure, while AHRQ estimates risk-adjustment parameters without shrinkage and then shrinks the resulting risk-adjusted rates in a second stage. There are two important potential advantages of a unified approach. The first is that the statistical uncertainty inherent in risk adjustment propagates naturally through to the final inference of interest. The second is that all model parameters are jointly estimated and their covariances are therefore appropriately accounted for. A practical potential disadvantage of a unified model is that parameter estimation depends on the full data set of raw rates across all hospitals, meaning that each time new data are to be considered, the full national-level model must be rerun. AHRQ's approach, by contrast, fixes these parameters a priori, making it possible for each hospital to estimate their own reliability-adjusted rates. In order to determine whether the statistical advantages of a unified approach suffice to outweigh its practical disadvantages, we recommend specifying and fitting a unified model to AHRQ's discharge-level dataset and comparing the resulting estimates to those from the current approach.

Evaluating methods for making inferences about differences in quality. We recommend further analysis of methods that fully incorporate statistical uncertainty in inferences based on hospital rates. In a Bayesian framework, these estimates of uncertainty could be used to enhance inferences when making hospital comparisons, such as the "exceedance probability" technique proposed by Ash et al. (2011). In assessing the performance of PSI rates in comparative reporting on patient safety, AHRQ could simulate hospital ranks based on exceedance probabilities and those based on the current or alternate modified approaches. In addition, we recommend that AHRQ study the possible role of stratification or peer grouping of hospitals by their characteristics as an approach to making inferences regarding comparative hospital performance.

REFERENCES

Ash, Arlene S., Stephen E. Fienberg, Thomas A. Louis, Sharon-Lise T. Normand, Therese A. Stukel, and Jessica Utts, with the Committee of Presidents of Statistical Societies. “Statistical Issues in Assessing Hospital Performance.” Baltimore, MD: Centers for Medicare & Medicaid Services, 2011, revised 2012.Bohl, Alex, David Jones, Sarah Schoenfeldt, and Dmitriy Poznyak. “Hitting the Target: Shrinking the AHRQ Quality Indicators to More Informed Priors.” Report submitted to the U.S Department of Health and Human Services, Agency for Healthcare Research and Quality. Cambridge, MA: Mathematica Policy Research, August 15, 2014.

Changes in Health Care Financing & Organization Initiative. “Issue Brief: Bridging Research and Policy to Advance Medicare’s Hospital Readmissions Reduction Program.” Washington, DC: HCFO, January 2014. Available at [http://www.hcfo.org/files/hcfo/HCFOPolicyBriefJanuary2014.pdf]. Accessed August 14, 2014.

Chen, Qi, Ann Borzecki, Amy Rosen, Hali Hambridge, Alex Bohl, Sheng Wang, Eric Schone, Frank Yoon, and David Jones. “Comparison of Standardization Methods for Risk Adjustment in the AHRQ Quality Indicators.” Report submitted to the U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality. Cambridge, MA: Mathematica Policy Research, August 18, 2014.

Dy, S.M., A. Zhang, C.M. Weston, B.D. Winters, R. Sharma, D. Jones, E. Schone, F. Yoon, A.K. Rosen, A. Borzecki, Q. Chen, and L.D. Engineer. “Improving the AHRQ Quality Indicators: Literature Review.” Washington, DC: Mathematica Policy Research, August 2013.

Fleiss, J.L., B. Levin, and M.C. Paik. *Statistical Methods for Rates and Proportions,* 3rd edition. Hoboken, NJ: John Wiley & Sons, Inc., 2003.

Friese, C.R., C.C. Earle, J.H. Silber, and L.H. Aiken. “Hospital Characteristics, Clinical Severity, and Outcomes for Surgical Oncology Patients.” *Surgery,* vol. 147, no. 5, 2010, pp. 602–609.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multi-level/hierarchical models*. New Yok: Cambridge University Press.

Ghosh, Malay, Georgios Papageorgiou and Janet Forresterb. “Multivariate Limited Translation Hierarchical Bayes Estimators.” *Journal of Multivariate Analysis*. Aug 2009; 100(7): 1398–1411.

Jones, David, Jessica Ross, Dmitriy Poznyak, Sam Stalley, and Alex Bohl. “Improving the AHRQ Quality Indicators: Exploratory Data Analyses of Differences in Quality Indicators by Hospital Characteristics.” Report submitted to the U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality. Cambridge, MA: Mathematica Policy Research, June 30, 2014a.

Jones, David, Dejene Ayele, Eric Schone, Sheng Wang, Alex Bohl, Frank Yoon, Xiaojing Lin, and Sarah Schoenfeldt. "The Effects of Incorporating Hospital Characteristics in Risk Adjustment for the AHRQ Quality Indicators." Report submitted to the U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality. Cambridge, MA: Mathematica Policy Research, August 15, 2014b.

Khuri, S.F., S.F. Najjar, J. Daley, B. Krasnicka, M. Hossain, W.G. Henderson, J.B. Aust, B. Bass, M.J. Bishop, J. Demakis, R. DePalma, P.J. Fabri, A. Fink, J. Gibbs, F. Grover, K. Hammermeister, G. McDonald, L. Neumayer, R.H. Roswell, J. Spencer, and R.H. Turnage. "VA National Surgical Quality Improvement Program. Comparison of Surgical Outcomes Between Teaching and Nonteaching Hospitals in the Department of Veterans Affairs." *Annals of Surgery,* vol. 234, no. 3, 2001, pp.382–383.

Kolfschoten, N.E., P.J. Marang van de Mheen, G.A. Gooiker, E.H. Eddes, J. Kievit, R.A. Tollenaar, M.W. Wouters, and the Dutch Surgical Colorectal Audit Group. "Variation in Case-Mix Between Hospitals Treating Colorectal Cancer Patients in the Netherlands." *European Journal of Surgical Oncology,* vol. 37, no. 11, 2011, pp. 956–963.

Medicare Payment Advisory Commission. "Refining the Hospital Readmissions Reduction Program." In *Report to the Congress: Medicare and the Health Care Delivery System.* Washington, DC: MedPAC, June 2013. Available at [http://www.medpac.gov/documents/reports/jun13_ch04.pdf?sfvrsn=0]. Accessed August 14, 2014.

Morris, C. "Parametric Empirical Bayes Inference: Theory and Application." *Journal of the American Statistical Association,* vol. 78, no. 381, 1983, pp. 47–55.

National Quality Forum. "Risk Adjustment for Sociodemographic Factors." Available at [http://www.qualityforum.org/risk_adjustment_ses.aspx]. Accessed August 14, 2014.

Silber, J.H., P.R. Rosenbaum, R.N. Ross, J.M. Ludwig, W. Wang, B.A. Niknam, P.A. Saynisch, O. Even-Shoshan, R.R. Kelz, and L.A. Fleisher. "A Hospital-Specific Template for Benchmarking its Cost and Quality." *Health Services Research,* vol. 49, no. 5, 2014, pp. 1475–97.

Wang, Sheng, Mariel Finucane, Xiaojing Lin, Frank Yoon, and David Jones. "Reliability Adjustment of the AHRQ Quality Indicators Under an Empirical Bayes Framework." Report submitted to the U.S Department of Health and Human Services, Agency for Healthcare Research and Quality. Cambridge, MA: Mathematica Policy Research, August 15, 2014.

APPENDIX A

SUMMARY OF PROPOSED ANALYSES TO EXAMINE FACTORS CONTRIBUTING TO DIFFERENCES IN HOSPITAL QI RATES BY HOSPITAL TYPE

This page has been left blank for double-sided copying.

The analyses below could be applied to improve the understanding of the factors that contribute to observed differences in hospital rates by hospital type.

- We could assess how large a difference in unmeasured risk would have to be by hospital type and how strong the correlation between the risk factor and the outcome would have to be for it to explain the observed difference in QI rates between two hospital types. That is, we would assess whether it is likely that any conceivable risk factor could explain the differences in QI rates.
- Similarly, we could remove a commonly accepted risk factor that is heavily correlated with hospital type to see how it affects differences by hospital type. A similar approach would be to simulate a risk factor effect that is the upper bound of observed risk factors or some reasonable multiplier of the effect (for example, double the effect). The objective is to determine how likely it is that the differences can be explained by unmeasured risk.
- Although there is no gold standard of hospital quality to which QI rates can be compared, it is generally accepted that measures of process are not as influenced by patient risk factors and can thus be measured more accurately. By comparing differences in the relationships of QI rates and hospital characteristics to the relationships of process measures and hospital characteristics, we could provide evidence regarding the likelihood that the former is driven by differences in quality rather than risk.
- A matched case-control approach that compares discharges within hospitals to discharges in other hospitals matched on an “enhanced” set of patient characteristics (for example, the primary payer or uncompensated care in addition to the risk factors in the current) could exert tighter control over the risk-adjustment process and provide evidence that any remaining differences are due to quality.
- Following the logic of a matched case-control study, we could apply a method called near/far matching, which incorporates an instrument that encourages the decision to seek care at one hospital type versus another, but is otherwise uncorrelated with the patient outcome. The outcomes of patients who are the same or similar according to all known risk factors, but who differ substantially with respect to the instrument are contrasted. This method provides a strong test of the impact of type by controlling for all known factors through matching, but ensuring the groups that are compared are very likely to use different hospital types. An example is how close the patient lives to hospitals of different types. QIs for patients who are otherwise similar but live at much different distances from a teaching hospital could be contrasted, to provide a measure of the impact of being “randomized” to treatment at a teaching hospital.
- As mentioned earlier in this report and highlighted in the EDA, the differences in QI rates by hospital characteristics could in part be indicative of the statistical properties of the QIs, namely the random variation that is inherent in the rates for small hospitals, which leads to mass points at extreme rates for small hospitals. Further analysis of these relationships could quantify the extent to which the extreme values are driving the differences by hospital type. For example, an analysis of how sample size influences the classification of hospital on CMS’s Hospital Compare site could help explain observed differences in the rates of outlier classification by hospital type.

- We recommend further analysis on the distribution of predicted risk scores to better understand how the models are predicting risk which is aggregated into expected rates; in particular, extreme values that do not appear to reflect the reality of risk posed by patients (at times indicating risk 1,000 times greater than that of the average patient). These extreme values can have a large impact given the rarity of many of the events targeted by the QIs.
- The broader set of risk-adjustment methods have also been critiqued for “insufficient” risk adjustment that understates the risk of complex patients, such as the risk for burn victims or patients suffering from multiple trauma. These risk factors are often partially or fully accounted for in the current risk-adjustment models. However, we recommend that enhancement or maintenance of the risk-adjustment models test the impact on comparisons by hospital type of adding burns or trauma as risk factors to all QI models.
- Further analysis on the difference in length of stay and present on admission coding by hospital type could add to the understanding of factors driving differences in QI rates.
- The resources required for hospitals to treat high-risk patients could lead to higher rates of adverse events at hospitals with high proportions of high-risk patients, all else being equal. If this is the case, even direct standardization approaches that match patients could still be missing a risk factor (an aggregate patient risk factor). For example, matching all of the high-risk patients at a hospital with a low proportion of high-risk patients to the high-risk patients at a hospital with a high proportion of high-risk patients could still miss the aggregate effect of the frequency of high-risk patients at the latter hospital. We recommend further study to assess whether such an aggregate patient risk factor exists and how it affects the QIs.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research