

## Appendix B: Detailed Methods

Empirical analyses were conducted to provide additional information about the indicators. These analyses were intended not as decision making tools, but rather explorations into the characteristics of the indicators. Specifically, these analyses explore the frequency and variation of the indicators, the potential bias, based on limited risk adjustment, and the relationship between indicators.

### Analysis Approach

*Data sources.* The data sources used in the empirical analyses were the 1997 Florida State Inpatient Database (SID) (for initial testing and development; 1995-1997 used for persistence analysis) and the 1997 State Inpatient Databases (SID) for 19 HCUP participating States, referred to in this report as the National SID (for the final empirical analysis). The Florida SID consists of about 2 million discharges from over 200 hospitals, and was chosen because Florida is a large diverse State. The National SID consists of about 19 million discharges from over 2,300 hospitals. The National SID contains all-payer data on hospital inpatient stays from participating States (Arizona, California, Colorado, Connecticut, Florida, Illinois, Iowa, Kansas, Maryland, Massachusetts, Missouri, New Jersey, New York, Oregon, Pennsylvania, South Carolina, Tennessee, Washington, and Wisconsin). All discharges from participating States' community hospitals are included in the SID database, which defines community hospitals as non-Federal, short-term, general, and other specialty hospitals, excluding long-term hospitals and hospital units of long-term care institutions, psychiatric hospitals, and alcoholism and chemical dependency treatment facilities.

A complete description of the content of the SID, including details of the participating States' discharge abstracts, can be found on the Agency for Healthcare Research and Quality Web site (<http://www.hcup-us.ahrq.gov/sidoverview.jsp>). Because the Florida SID was used only for initial testing and development, the empirical results reported are from the National SID. Descriptive results from the Florida SID are reported for comparison to ensure that the hospital-level results were similar in both data sources. Differences between Florida and national results are pointed out in the text. The National SID data were also used for the construction of area measures, with data from the U.S. Census Bureau used to construct the denominator of these rates.

*Reported patient safety indicators.* Three sets of patient safety indicators were examined. First, the Accepted patient safety indicators met the face validity criteria established through the literature review and clinician panel review. Second, the Experimental patient safety indicators did not meet those criteria, but appeared to warrant further testing and evaluation. Third, several Accepted patient safety indicators were modified into *area* indicators, which were designed to assess the total incidence of the adverse event within geographic areas. For example, the project team constructed an indicator for "Transfusion reaction" at both the hospital and area levels. Transfusion reactions that occur after discharge from a hospitalization would result in a readmission. The area-level indicator includes these cases, while the provider level restricts the number of transfusion reactions to only those that occur during the same hospitalization that exposed the patient to this risk.

All potential indicators were examined empirically by developing and conducting statistical tests for precision, bias, and relatedness of indicators. For each indicator, the project team calculated five different estimates of provider level performance:

1. The raw indicator rate was calculated using the number of adverse events in the numerator divided by the number of discharges in the population at risk by hospital. For the area indicators, the denominator is the population of the Metropolitan Statistical Area (MSA), New England County Metropolitan Area (for the New England States) or county (for non-MSA areas) of the hospital.
2. The raw indicator was adjusted using a logistic regression to account for differences among hospitals (and areas) in demographics (specifically, age and gender). Age was modeled using a set of dummy variables to represent 10-year categories except for young children, whose age categories are

narrower (i.e., less than 1, 1-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, and 85 or more years), along with a parallel set of age-gender interactions. Because of sparse cells, certain age categories were combined or omitted for selected indicators, such as the obstetric indicators.

3. The raw indicator was adjusted to account for differences among hospitals in age, gender and modified DRG category (as described below).
4. The raw indicator was adjusted to account for differences among hospitals in age, gender, modified DRG, and comorbidities (defined using an adaptation of the AHRQ comorbidity software) of patients.
5. Multivariate signal extraction (MSX) methods were applied to adjust for reliability by estimating the amount of “noise” (i.e., variation due to random error) relative to the amount of “signal” (i.e., systematic variation in hospital performance or the ‘reliability’) for each indicator. This or similar “reliability adjustment” has been used in the literature for similar purposes.<sup>1 2</sup> Multivariate methods (taking into account correlations among indicators to extract additional signal) were applied to most of the accepted indicators. The exceptions were Death in Low Mortality DRGs and Failure to Rescue. Only univariate signal extraction methods (smoothing) were applied to these two indicators and to the experimental indicators, because these indicators possibly cover broader clinical concepts. Correlations between these indicators and other indicators may not reflect correlations due to quality of care, and thus inclusion of these indicators may adversely affect the MSX approximations.

For additional details on the empirical methods, refer to the companion EPC HCUP Quality Indicator Report, published by AHRQ (<http://www.qualityindicators.ahrq.gov/downloads.htm>). Additional details on the modifications made to the DRG and comorbidity categories are described below.

*Hospital Fixed Effects.* In the risk-adjustment models, hospital fixed effects were calculated using the standard method with logistic models of first estimating the predicted value for each discharge, then subtracting the actual outcome from the predicted, and averaging the difference for each hospital to get the hospital fixed effect estimate. In the Quality Indicator Report,<sup>3</sup> linear regression models were used with hospital fixed effects included, arguing that the logistic approach yielded biased estimates due to the omission of a variable (the hospital) correlated with both the dependent (e.g., in-hospital mortality) and the independent (e.g., age, gender, APR-DRG) variables in the model. Given the rare occurrence of many of the PSIs, however, the logistic approach may be more appropriate for this application. Linear methods assume that the error term is normally distributed. This assumption is violated when the outcome is dichotomous.

The QI means were generally an order of magnitude higher than the PSI means, so the assumption was not as problematic. However, the most appropriate method depends on the particular characteristics of each indicator, whether QI or PSI. To the extent that bias is a concern, accounting for the clustering of patients by using a hospital fixed effect is advantageous. To the extent that extreme values are a concern, imposing structure on the error term with logistic methods is advantageous. In the end, the two approaches can be compared in terms of how much difference it makes in the relative assessment of provider performance. This issue warrants further analysis to better understand the trade-offs and limitations of each approach, and under what conditions and for what indicators each approach might best apply.

Specifically, the risk-adjusted “raw” estimate of a hospital’s performance is constructed in two steps. In the first step, if it is denoted whether or not the event associated with a particular indicator  $Y^k$  ( $k=1, \dots, K$ ) was observed for a particular patient  $i$  in year  $t$  ( $t=1, \dots, T$ ), then the regression to construct a risk-adjusted “raw” estimate of a particular patient’s performance on each indicator can be written as:

$$(1) \quad Y_{it}^k = Z_{it} \Pi_t^k + \xi_{it}^k, \quad \text{where}$$

<sup>1</sup> Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician “report cards” for assessing the costs and quality of care of a chronic disease JAMA 1999;28(22):2098-105.

<sup>2</sup> Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. Ann Intern Med 1997;127(8 Pt 2):764-8).

<sup>3</sup> Davis et al. 2001.

$Y_{it}^k$  is the  $k^{\text{th}}$  PSI for patient  $i$  in year  $t$  (i.e., whether or not the event associated with the indicator occurred on that discharge).

$Z_{it}$  is a vector of patient covariates for patient  $i$  in year  $t$  (i.e., the patient-level measures used as risk adjusters).

$\Pi_t^k$  is a vector of parameters in each year  $t$ , giving the effect of each patient risk adjuster on indicator  $k$  (i.e., the magnitude of the risk adjustment associated with each patient measure).

$\varepsilon_{it}^k$  is the unexplained residual in this patient-level model.

In the second step, the hospital effect was estimated by subtracting the resulting predictions from this patient-level regression from the actual observed patient-level outcomes, and taking the mean of this difference for each hospital. That is, for each hospital  $j$  ( $j=1, \dots, J$ ),

$$(2) \quad M_{jt}^k = Y_{ijt}^k - (Z_{it} \Pi_t^k + \xi_{it}^k), \quad \text{where}$$

$M_{jt}^k$  is the “raw” adjusted measure for indicator  $k$  for hospital  $j$  in year  $t$  (i.e., the hospital “fixed effect” in the patient-level regression).

$Z_{it}$  is the vector of patient covariates for patient  $i$  in year  $t$  estimated in Step 1.

In addition to age, sex, and age\*sex interactions as adjusters in the model, the project team also included a modified DRG and comorbidity category for the admission.

*Modified DRG Categories.* Two modifications were made to the Centers for Medicare and Medicaid Services (CMS, formerly Health Care Financing Administration) DRGs. First, adjacent DRG categories that were separated by the presence or absence of comorbidities or complications were collapsed. For example, DRGs 076 (Other Resp System Operating Room Procedures w CC) and 077 (Other Resp System Operating Room Procedures w/o CC) were grouped into one category. The purpose was to avoid adjusting for the complication the team was trying to measure. Second, most of the super-MDC DRG categories were excluded from the logistic models. Excluding these categories also avoids adjusting for the complications the team was trying to measure. For example, tracheostomies (DRG 482-483) often result from potentially preventable respiratory complications that require long-term mechanical ventilation. Similarly, operating room procedures unrelated to the principal diagnosis (DRG 468, 477) often result from potentially preventable complications that require surgical repair (i.e., fractures, lacerations).

In the companion technical report on quality indicators, the risk adjustment method implemented All Patient Refined (APR)-DRGs, a refinement of DRGs to capture different levels of complications. However, patient safety indicators, designed to detect potentially preventable complications, require a risk adjustment approach that does not inherently remove the differences between patients based on their complications. The APR-DRGs could be modified to remove applicable complications, on an indicator-by-indicator basis, but implementation of such an approach was beyond the scope of the current project. In this report, APR-DRG risk adjustment was not implemented.

*Modified Comorbidity Software.* To adjust for comorbidities, the project team used an updated adaptation of AHRQ Comorbidity Software (<http://www.hcup-us.ahrq.gov/toolsssoftware/comorbidity/comorbidity.jsp>). The ICD-9-CM codes used to define the comorbidity categories were modified to address four main issues.

1. Comorbidity categories were excluded in the current software that include conditions likely to represent potentially preventable complications in certain settings, such as after elective surgery. Specifically, three DRG categories (cardiac arrhythmia, coagulopathy, and fluid/electrolyte disorders) were removed from the comorbidity adjustment.

2. Most adaptations were designed to capture acute sequelae of chronic comorbidities, where both conditions are represented by a single ICD-9-CM code. For example, the definition of hypertension was broadened to include malignant hypertension, which usually arises in the setting of chronic hypertension. Unless these "acute on chronic" comorbidities are captured, some patients with especially severe comorbidities would be mislabeled as not having conditions of interest.
3. The comorbidity definitions did not include obstetric comorbidity codes, which are relevant for the obstetric indicators. Codes, when available, for these comorbidities in obstetric patients were added.
4. Slight updating was necessary based on recent ICD-9-CM code changes.

*Low Mortality DRGs.* In order to be included in the "Low Mortality DRG" indicator, the DRG had to have an overall in-hospital mortality rate (based on the National SID sample) of less than 0.5%. In addition, if a DRG category was split based on the presence of comorbidities or complications, then the category was included only if both DRGs (with and without comorbidities or complications) met the mortality threshold. Otherwise, the category was not included in the "Low mortality DRG" PSI. The indicator is reported as a single measure and stratified into medical (adult and pediatric), surgical (adult and pediatric), neonatal, obstetric and psychiatric DRGs.

## **Empirical Analysis Statistics**

Using these methods, the project team constructed a set of statistical tests to examine precision, bias, and relatedness of indicators for all accepted hospital-level indicators, and precision and bias for all accepted area-level and experimental indicators. Each of the key statistical test results was summarized and explained in the overview section of the companion HCUP Quality Indicator report.<sup>4</sup> Tables B-1 through B-3 provide a summary of the statistical analyses and their interpretation.

---

<sup>4</sup> Davies et al., 2001.

**Table B-1. Precision Tests**

Measure	Statistic/Adjustments		Interpretation
<b>Precision. Is most of the variation in an indicator at the level of the hospital? Do smoothed estimates of quality lead to more precise measures?</b>			
a. Observed variation in indicator	Hospital-Level Standard Deviation Hospital -Level Skew Statistic	Unadjusted Age-gender adjusted Modified DRG adjusted Modified AHRQ comorbidity adjusted	Risk adjustment can either increase or decrease observed variation. If increase, then differences in patient characteristics mask provider differences. If decrease, then differences in patient characteristics account for provider differences.
b. MSX methods	Signal Standard Deviation Signal Share Signal Ratio	Reliability adjusted	Estimates what percentage of the observed variation between hospitals reflects systematic differences versus random noise. Signal share is a measure of how much of the total variation (patient and provider) is potentially subject to hospital control.

**Table B-2. Bias Tests**

Measure	Statistic	Interpretation
<b>Bias. Does risk adjustment change our assessment of relative hospital performance, after accounting for reliability? Is the impact greatest among the best or worst performers, or overall? What is the magnitude of the change in performance?</b>		
MSX methods: unadjusted vs. age, sex, modified DRG, comorbidity risk adjustment	Spearman Rank Correlation Coefficient (before and after risk adjustment)	Risk adjustment matters to the extent that it alters the assessment of relative hospital performance. This test determines the impact overall.
	Average absolute value of change relative to mean (after risk adjustment)	This test determines whether the absolute change in performance was large or small relative to the overall mean.
	Percentage of the top 10% of hospitals that remains the same (after risk adjustment)	This test measures the impact at the highest rates (in general, the worse performers).
	Percentage of the bottom 10% of hospitals that remains the same (after risk adjustment)	This test measures the impact at the lowest rates (in general, the better performers).
	Percentage of hospitals that move more than two deciles in rank (up or down) (after risk adjustment)	This test determines the magnitude of the relative changes.

**Table B-3. Relatedness Tests**

Measure	Statistic	Interpretation
<b>Relatedness of indicators. Is the indicator related to other indicators in a way that makes clinical sense? Do methods that remove noise and bias make the relationship clearer?</b>		
a. Correlation of indicator with other indicators	Spearman correlation coefficient	Are indicators correlated with other indicators in the direction one might expect?
b. Factor loadings of indicator	Factor loadings, based on Spearman correlation, Principal Component Analysis	Do indicators load on factors with other indicators that one might expect?