

NATIONAL QUALITY FORUM—Composite Measure Testing (subcriteria 2a2, 2b2-2b7, 2d)

Measure Number (if previously endorsed): **532**

Composite Measure Title: [Pediatric Patient Safety for Selected Indicators \(PDI #19\)](#)

Date of Submission: [Click here to enter a date](#)

Composite Construction:

☒ Two or more individual performance measure scores combined into one score

☐ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)

☐ Any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient)

Instructions: Please contact NQF staff before you begin.

- If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure submission.
- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- For all composite measures, sections 1, 2a2, 2b2, 2b3, 2b5, and 2d must be completed.
- For composites with outcome and resource use measures, section 2b4 also must be completed.
- If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), section 2b6 also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2), validity (2b2-2b6), and composites (2d) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*including questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](#).

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing ¹⁰ demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing ¹¹ demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; ¹²

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient

preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). ¹³

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; ^{14,15} and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** ¹⁶ **differences in performance;**

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:

2d1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and

2d2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.

(if not conducted or results not adequate, justification must be submitted and accepted)

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of \$25 in cost for an episode of care (e.g., \$5,000 v. \$5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing, (e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation. If different data sources are used for different components in the composite, indicate the component after the checkbox.)

Measure Specified to Use Data From: (must be consistent with data sources entered in S.23)	Measure Tested with Data From:
<input type="checkbox"/> abstracted from paper record	<input type="checkbox"/> abstracted from paper record
<input checked="" type="checkbox"/> administrative claims	<input checked="" type="checkbox"/> administrative claims
<input type="checkbox"/> clinical database/registry	<input type="checkbox"/> clinical database/registry
<input type="checkbox"/> abstracted from electronic health record	<input type="checkbox"/> abstracted from electronic health record
<input type="checkbox"/> eMeasure (HQMF) implemented in EHRs	<input type="checkbox"/> eMeasure (HQMF) implemented in EHRs
<input type="checkbox"/> other: Click here to describe	<input type="checkbox"/> other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).

All analyses were completed using data from the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID), 2007-2011. HCUP is a family of health care databases and related software tools and products developed through a Federal-State-Industry partnership and sponsored by the Agency for Healthcare Research and Quality (AHRQ). HCUP databases bring together the data collection efforts of State data organizations, hospital associations, private data organizations, and the Federal government to create a national information resource of encounter-level health care data. The HCUP SID contain the universe of the inpatient discharge abstracts in participating States, translated into a uniform format to facilitate multi-State comparisons and analyses. Together, the SID encompass about 97 percent of all U.S. community hospital discharges (in 2011, 46 states participated for a total of more than 38.5 million hospital discharges; of which approximately 5 million hospital discharges were for children 17 years and younger [inclusive of uncomplicated births]). As defined by the American Hospital Association, community hospitals are all non-Federal, short-term, general or other specialty hospitals, excluding hospital units of institutions. Veterans hospitals and other Federal facilities are excluded. Children's general and specialty hospitals are included in the universe of hospitals. Taken from the Uniform Bill-04 (UB-04), the SID data elements include ICD-9-CM coded principal and secondary diagnoses and procedures, additional detailed clinical and service information based on revenue codes, admission and discharge status, patient demographics, expected payment source (Medicare, Medicaid, private insurance as well as the uninsured), total charges and length of stay (www.hcup-us.ahrq.gov)

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP).. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 4.5)

1.3. What are the dates of the data used in testing? 2007-2011

1.4. What levels of analysis were tested? *(testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan)*

Measure Specified to Measure Performance of: (must be consistent with levels entered in item S.26)	Measure Tested at Level of:
<input type="checkbox"/> individual clinician	<input type="checkbox"/> individual clinician
<input type="checkbox"/> group/practice	<input type="checkbox"/> group/practice
<input checked="" type="checkbox"/> hospital/facility/agency	<input checked="" type="checkbox"/> hospital/facility/agency
<input type="checkbox"/> health plan	<input type="checkbox"/> health plan
<input type="checkbox"/> other: Click here to describe	<input type="checkbox"/> other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? *(identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample)*

The hospital universe is defined as all hospitals located in the U.S. that are open during any part of the calendar year and designated as community hospitals in the AHA Annual Survey Database (Health Forum, LLC © 2011). The AHA defines community hospitals as follows: "All non-Federal, short-term, general, and other specialty hospitals, excluding hospital units of institutions." Starting in 2005, the AHA included long term acute care facilities in the definition of community hospitals. These facilities provide acute care services to patients who need long term hospitalization (stays of more than 25 days). Consequently, Veterans Hospitals and other Federal facilities (Department of Defense and Indian Health Service) are excluded. Beginning in 1998, we excluded short-term rehabilitation hospitals from the universe because the type of care provided and the characteristics of the discharges from these facilities were markedly different from other short-term hospitals. General and specialty children's hospitals are included in the hospital universe.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample)

Table 1. Reference Population

Year/ Characteristic	Hospitals	Outcome of Interest	Population at Risk	Overall Composite Performance Score
2011	4,594	-	1,068,839	1.000
2010	4,603	-	1,082,230	1.000
2009	4,532	-	1,116,717	1.000
2008	4,496	-	1,093,153	1.000
2007	4,264	-	1,025,900	1.000
Composite Performance Score Distribution 2011				
	5th	25th	Median	75th
	0.289	0.590	0.898	1.300
				2.059

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2007-2011. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 4.5)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Not applicable

2a2. RELIABILITY TESTING

2a2.1. What level of reliability testing was conducted?

Note: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.

☐ **Performance measure score** (e.g., signal-to-noise analysis)

2a2.2. Describe the method of reliability testing and what it tests (describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used)

Our metric of reliability is the signal to noise ratio, which is the ratio of the between hospital variance (signal) to the within hospital variance (noise). The formula is $\text{signal} / (\text{signal} + \text{noise})$. There is hospital-specific signal to noise ratio, which is used as an Empirical Bayes univariate shrinkage estimator. The overall signal to noise ratio is a weighted average of the hospital-specific signal-to-noise ratio, where the weight is $[1 / (\text{signal} + \text{noise})^2]$. The signal is calculated using an iterative method. The analysis reports the reliability of the risk-adjusted rate (before applying the empirical Bayes univariate shrinkage estimator).

2a2.3. What were the statistical results from reliability testing? (e.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis)

Note: The provider level composite is a weighted average of reliability- and risk-adjusted component ratios; no reliability metrics are calculated for the composite. Reported are the reliability metrics for the component measures.

Table 2. Reliability by Component Measure

Component	Number of Hospitals	Ave. Number of Patients per Hospital	Ave. Signal-to-Noise Ratio for Hospitals	Percent of Signal Variance Explained by Performance Score
PDI 01	4,699	651.5	0.71820	0.54979
PDI 02	3,347	116.0	0.77829	0.64653
PDI 05	4,690	592.3	0.55336	0.63944
PDI 10	1,972	45.8	0.76229	0.79419
PDI 11	2,520	23.9	0.71705	0.85944
PDI 12	4,594	528.1	0.75911	0.75209

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 4.5)

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.e., what do the results mean and what are the norms for the test conducted?)

The composite is a numerator weighted average of reliability adjusted component ratios. Therefore the component measures that have greater reliability contribute more to the composite performance score. Eventhough the PDI events are infrequent, the large denominators generally means that the average reliability across all hospitals (patient weighted) is moderate to high.

2b2. VALIDITY TESTING

Note: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.

2b2.1. What level of validity testing was conducted?

☒ **Composite performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance)

☐ **Systematic assessment of content validity**

☒ **Validity testing for component measures** (check all that apply)

Note: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.

☒ **Endorsed (or submitted) as individual performance measures**

☐ **Critical data elements** (data element validity must address ALL critical data elements)

- ☐ Empirical validity testing of the component measure score(s)
- ☐ Systematic assessment of face validity of component measure score(s) as an indicator of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used*)

We conduct construct validity testing to examine the association between the composite performance score and hospital structural characteristics potentially associated with quality of care, including prior performance, using regression analysis.

Table 3. Structure Measures Used to Estimate Prior Probability

Measure	How it is measured	Rationale
Ln(Volume)	Natural log of the denominator	Practice makes perfect or referral
Reservation Quality	Inverse of average daily census (ADC)	Reflects the excess capacity in the inputs of production (e.g. nurse staffing)
Transfer Out	Overall percent transfer out	Routine transferring of particular categories of patients
Maximum DX	Maximum reported diagnosis codes	Higher prevalence and co-morbidities
Prior Performance	Prior year composite performance score	Share of performance likely to persist

The hypothesized relationship is as follows:

- Volume: Higher volume is associated with better outcomes, either because practice makes perfect (volume causes outcome) or referral (outcome causes volume)
- Reservation quality: Higher reservation quality is associated with better outcomes because reservation quality is associated with excess capacity
- Transfer out: Higher transfer out rate is associated with better outcomes because transferred cases have higher risk of mortality or adverse outcome
- Diagnosis codes: More reported diagnosis codes are associated with more reported comorbidities, therefore higher expected rates, therefore better outcomes

2b2.3. What were the statistical results from validity testing? (e.g., correlation; t-test)

Table 4. Regression on Structure Measures

Variable	Label	Coef.	Std. Err.	t	P> t	[95% Conf.	Interval]
Invol	Ln(Volume)	0.122865	0.012260	10.02	0.0000	0.09883	0.14690
adcin	Reservation Quality	-0.000137	0.000747	-0.18	0.8550	-0.00160	0.00133
trnsout	Transfer Out	-1.135826	0.310066	-3.66	0.0000	-1.74370	-0.52795
maxdx	Maximum DX	0.009944	0.002045	4.86	0.0000	0.00593	0.01395
_cons	Constant	-0.507002	0.075632	-6.70	0.0000	-0.65528	-0.35873
Invol	Ln(Volume)	0.050833	0.011181	4.55	0.0000	0.02891	0.07275
adcin	Reservation Quality	-0.000255	0.000407	-0.63	0.5310	-0.00105	0.00054
trnsout	Transfer Out	-0.716129	0.153400	-4.67	0.0000	-1.01687	-0.41539
maxdx	Maximum DX	-0.001151	0.001016	-1.13	0.2570	-0.00314	0.00084
prior2	Prior Performance	0.695913	0.018140	38.36	0.0000	0.66035	0.73148
_cons	Constant	-0.152698	0.053610	-2.85	0.0040	-0.25780	-0.04760

Note: the dependent variable in the regression is the composite performance score

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.e., what do the results mean and what are the norms for the test conducted?)

Hospitals with higher volume have worse performance (higher ratio) and hospitals with a higher transfer out rate have better performance (lower ratio). Hospitals that report on average more diagnosis codes have worse performance (higher ratio). Conditional on prior performance, hospitals with higher volumes have worse or better performance (that is, current volume provides new information) and hospitals with higher transfer out rates have better performance (that is, current transfer out rate provides new information). Overall performance is moderately persistent over time.

2b3. EXCLUSIONS ANALYSIS

Note: Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.

NA ☐ no exclusions — skip to section [2b4](#)

2b3.1. Describe the method of testing exclusions and what it tests (describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used)

2b3.2. What were the statistical results from testing exclusions? (include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (i.e., the value outweighs the burden of increased data collection and analysis. **Note:** If patient preference is an exclusion, the measure must be specified

so that the effect on the performance score is transparent, e.g., scores with and without exclusion)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

Note: Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section [2b5](#).

2b4.1. What method of controlling for differences in case mix is used? (check all that apply)

☐ Endorsed (or submitted) as individual performance measures

☒ No risk adjustment or stratification

☐ Statistical risk model

☐ Stratification by risk categories

☐ Other, [Click here to enter description](#)

2b4.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

Note: The provider level composite is a weight average of reliability- and risk-adjusted component ratios; no discrimination or calibration metrics are calculated for the composite itself

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of $p < 0.10$; correlation of x or higher; patient factors should be present at the start of care and not related to disparities)

Not applicable

2b4.4. What were the statistical results of the analyses used to select risk factors?

Not applicable

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (describe the steps—do not just name a method; what statistical analysis was used)

Not applicable

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

if stratified, skip to [2b4.9](#)

2b4.6. Statistical Risk Model Discrimination Statistics (e.g., c-statistic, R-squared):

Not applicable

2b4.7. Statistical Risk Model Calibration Statistics (e.g., Hosmer-Lemeshow statistic):

Not applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Not applicable

2b4.9. Results of Risk Stratification Analysis:

Not applicable

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.e., what do the results mean and what are the norms for the test conducted?)

Not applicable

***2b4.11. Optional Additional Testing for Risk Adjustment (*not required*, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed)**

Not applicable

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

Note: *Applies to the composite performance measure.*

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)

We calculate the posterior probability distribution for each hospital parameterized using the Gamma distribution. We then calculate the probability that the hospital is better or worse than the reference population benchmark (20th percentile) or threshold (80th percentile) composite performance score at a 95 percent probability overall and by hospital size decile. The analysis is with the computed composite performance scores for the measure as specified (including shrinkage estimator).

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined)

Table 5. Performance Categories by Hospital Size Decile

Size Decile	Number of Hospitals	Ave. Number of patients per Hospital in Decile	Benchmark		Threshold	
			Proportion Better	Proportion Worse	Proportion Better	Proportion Worse
1	460	3.9	0.24130	0.09348	0.28043	0.03043
2	459	12.5	0.66231	0.01089	0.70370	0.00218
3	460	37.1	0.88913	0.00217	0.89348	0.00000
4	459	107.5	0.93246	0.00000	0.93246	0.00000
5	459	248.8	0.94553	0.00000	0.95861	0.00000
6	460	465.3	0.85217	0.00000	0.90870	0.00000
7	459	825.6	0.61002	0.00436	0.76906	0.00000
8	460	1,407.2	0.16739	0.00652	0.44783	0.00217
9	459	2,471.7	0.03486	0.01961	0.13290	0.00218
10	459	8,387.8	0.01961	0.19390	0.08061	0.06100
	4,594	1,396.0	0.53548	0.03309	0.61080	0.00980
Patient weighted			0.12383	0.18340	0.22789	0.06259

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 4.5)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.e., what do the results mean in terms of statistical and meaningful differences?)

Small hospitals are more likely to perform better than the benchmark and large hospitals are less likely to perform better than the benchmark. Very small or very large hospitals are more likely to perform worse than the threshold.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

Note: Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.

If only one set of specifications for each component, this section can be skipped.

Note: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). **If comparability is not demonstrated, the different specifications should be submitted as separate measures.**

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications (describe the steps—do not just name a method; what statistical analysis was used)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (e.g., correlation, rank order)

Not applicable

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.e., what do the results mean and what are the norms for the test conducted?)

Not applicable

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Note: *Applies to the overall composite measure.*

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias *(describe the steps—do not just name a method; what statistical analysis was used)*

Not applicable

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each)

Not applicable

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias? *(i.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data)*

Not applicable

2d. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH

Note: *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

2d1.1 Describe the method used *(describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

The quality construct is that use of the composite by consumers making selection decisions or providers allocating resources for change is less likely to result in wasted effort. Our method is to conduct a correlation analysis to ensure that worse performance on the composite is associated with worse performance on the component measures.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., correlations, contribution of each component to the composite score, etc.; if no empirical analysis, identify the components that were considered and the pros and cons of each)

	PDI 01	PDI 02	PDI 05	PDI 10	PDI 11	PDI 12
PDI 01	1.0000					
PDI 02	-0.0242	1.0000				
PDI 05	0.2390	0.0325	1.0000			
PDI 10	0.0928	0.0625	0.0754	1.0000		
PDI 11	-0.0153	0.0718	0.1083	0.1655	1.0000	
PDI 12	0.0876	0.0820	0.2716	0.1850	0.0999	1.0000

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i.e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)

At the hospital level, the component measures are positively correlated with each other, with the exception of PDI 01, which is an infrequent event. Therefore use of the composite does not require trade-offs among component measures.

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible

2d2.1 Describe the method used (describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)

The composite is a weighted average of reliability-adjusted observed to expected ratios, where the component weights are the relative frequency of the numerator in the reference population. The concept is the use of the composite minimizes the likelihood of harm associated with a potentially preventable adverse event where that likelihood is expressed as the probability of an potentially preventable adverse event x harm association with the event (in the current specification all events are assigned equal harm). The rationale is that numerator weights reflect the probability that an individual patient would experience a particular adverse event.

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

Table 14. NQF Numerator Weights for PDI 19

Indicator	Weight ¹	Ave. Signal-to-Noise Ratio for Hospitals	Correlation With Composite
PDI 01 Accidental Puncture or Laceration Rate	0.3119	0.71820	0.0876
PDI 02 Pressure Ulcer Rate	0.0100	0.77829	0.0820
PDI 05 Iatrogenic Pneumothorax Rate	0.0701	0.55336	0.2716
PDI 010 Postoperative Sepsis Rate	0.2655	0.76229	0.1850
PDI 011 Postoperative Wound Dehiscence Rate	0.0121	0.71705	0.0999
PDI 012 Central Venous Catheter-Related Blood Stream Infection Rate	0.3304	0.75911	1.0000
SUM	1.0000		

¹ Based on the use of present on admission (POA) data (i.e. USEPOA = 1). Indicators with a weight of zero are not included in the composite calculation for Version 4.5.

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.e., *what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting*)

By construction, adverse events that are less common and less reliable contribute less to the composite performance score. Performance on the composite is most highly associated with performance on PDI 05 and PDI 12, followed by PDI 10. PDI 01, PDI 02 and PDI 11 contribute less.