

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

Measure Number (if previously endorsed): 0272

Measure Title: Diabetes Short-Term Complications Admission Rate (PQI 01)

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

Type of Measure:

<input type="checkbox"/> Composite – STOP – use composite testing form	<input checked="" type="checkbox"/> Outcome (including PRO-PM)
<input type="checkbox"/> Cost/resource	<input type="checkbox"/> Process
<input type="checkbox"/> Efficiency	<input type="checkbox"/> Structure

Instructions

- 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 measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- For 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) and validity (2b2-2b6) 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 20 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.

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 the numerator and denominator, indicate N [numerator] or D [denominator] 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 checked="" type="checkbox"/> other: US Census	<input checked="" type="checkbox"/> other: US Census

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). 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. 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).

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)

The area universe is defined as the county of the residence of the patient for discharges in the hospital universe. 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.

Population estimates are derived from the US Census and are detailed in the 2013 Population File for Use with the AHRQ Quality Indicators posted on the AHRQ QI website: <http://www.qualityindicators.ahrq.gov/Downloads/Software/SAS/V45/AHRQ%20QI%20Population%20File%20V4.5.pdf> and provided in the supplemental materials. Public-use files of intercensal and postcensal estimates of county-level population by five-year age group, sex, race, and Hispanic origin were acquired from the Census Bureau (<http://www.census.gov/popest/>) covering the years 1995 through 2011.

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 type="checkbox"/> hospital/facility/agency	<input type="checkbox"/> hospital/facility/agency
<input type="checkbox"/> health plan	<input type="checkbox"/> health plan
<input checked="" type="checkbox"/> other: Geographic area (county)	<input checked="" type="checkbox"/> other: Geographic area (county)

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)

Table 1. Reference Population

Year	Nbr Counties (Areas)	Outcome of Interest (Numerator)	Population at Risk (Denominator)	Observed Rate Per 100,000
2011	3,112	153,410	236,853,390	64.77
2010	3,111	142,019	234,354,341	60.60
2009	3,112	131,916	231,837,944	56.90
2008	3,111	125,493	229,336,422	54.72
2007	3,107	112,232	226,778,104	49.49

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)

1.6. 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)*

See 1.5 (Table 1)

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

Note: *If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? *(may be one or both levels)*

☐ **Critical data elements used in the measure** *(e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements)*

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

2a2.2. For each level checked above, 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 county (area) variance (signal) to the within area variance (noise). The formula is $\text{signal} / (\text{signal} + \text{noise})$. There is an area-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 county (area)-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. For each level of testing checked above, 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)

Table 2. Reliability by County (Area) Size Decile

Size Decile	Number of Areas	Ave. Number of Persons per Area in Decile	Ave. Signal-to-Noise Ratio for Areas in Decile	Percent of Signal Variance Explained by Performance Score
1	312	2,278.3	0.48026	0.79120
2	311	5,657.8	0.72456	0.84477
3	311	8,817.1	0.80578	0.87430
4	311	12,641.0	0.85739	0.89810
5	311	17,289.0	0.89213	0.91718
6	312	23,989.3	0.92000	0.93469
7	311	33,768.0	0.94222	0.95030
8	311	53,199.9	0.96215	0.96583
9	311	103,760.9	0.97998	0.98108
10	311	500,100.9	0.99348	0.99363
Overall	3,112	76,109.7	0.97207	0.98166

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?)

Overall the risk-adjusted rate is highly reliable. Based on a norm of a signal-to-noise ratio of 0.80, 80% of counties (areas) exceed the norm. Reliability is less than the norm in counties (areas) with population less than approximately 6,000 persons, meaning that the performance score is reliability adjusted closer to the shrinkage target in those counties (areas).

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (may be one or both levels)

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

☒ 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)

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 risk-adjusted rate and area 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	Less Access to High Quality Outpatient Care Construct (F1)	Less Market Competition Construct (F2)
MD Density	Number of Physicians in Patient Care per Person	Areas with less physicians per person have less access to high quality outpatient care	Areas with more physicians per person have less market competition
Excess Capacity	Percent of Available Short-term General Hospital Beds per Total Beds	Areas with greater excess bed capacity have supply side incentive to have greater rates of admission	Areas with less excess bed capacity have less market competition
Poverty Status	Percent of Persons in Poverty	Areas with greater persons in poverty have less access to high quality outpatient care	Areas with greater persons in poverty have less market competition
Insurance Status	Percent of Persons (Under 65) without Health Insurance	Areas with greater persons without health insurance have less access to high quality outpatient care	Areas with greater persons without health insurance have less market competition
Population Density	Population Density per Square Mile	Areas with less population density have less access to high quality outpatient care	Areas with more population density have less market competition

Source: Area Health Resource File (ARF) 2012-2013. US Department of Health and Human Services, Health Resources and Services Administration, Bureau of Health Professions, Rockville, MD.

NOTE: Areas defined as counties in the ARF and analyses presented here.

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]
F1	Access to Quality Care	0.000132	0.000013	10.54	0.0000	0.00011 0.00016
F2	Market Competition	0.000154	0.000021	7.26	0.0000	0.00011 0.00020
_cons	Constant	0.000697	0.000009	80.74	0.0000	0.00068 0.00071
F1	Access to Quality Care	0.000025	0.000006	3.92	0.0000	0.000013 0.000038
F2	Market Competition	0.000035	0.000009	3.89	0.0000	0.000017 0.000053
prior2	Prior Performance	0.909917	0.036917	24.65	0.0000	0.837534 0.982300
_cons	Constant	0.000147	0.000021	6.92	0.0000	0.000105 0.000188

Note: the dependent variable in the regression is the risk adjusted rate

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)

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?)

Given the stated rationale, the expectation for the regression analysis given the expected relationship between the “Less Access to High Quality Outpatient Care” construct validity measure (F1) and the county (area) risk-adjusted rate is a positive, statistically significant coefficient. The expectation for the regression analysis given the expected relationship between the “More Market Competition” construct validity measure (F2) and the county (area) risk-adjusted rate is a positive, statistically significant coefficient. The results are consistent with expectations. Also, past performance is a strong predictor of current performance with a coefficient of 0.91.

2b3. EXCLUSIONS ANALYSIS

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)

Not applicable

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)

Not applicable

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*)

Not applicable

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

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?

- ☐ No risk adjustment or stratification
- ☒ Statistical risk model with risk factors
- ☐ Stratification by [Click here to enter number of categories](#) risk categories
- ☐ Other, [Click here to enter description](#)

2b4.2. If an outcome or resource use 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.

Not applicable

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)

For the area level indicators, the models use the complete set of covariates for gender, age in 5-year age groups, an interaction with gender * age. There is also an optional set of covariates for poverty category based on the county of patient residence.

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

The process to select risk factors is described in the AHRQ QI Empirical Methods report. The results of the analyses are provided in the PQI Parameter Estimates document. Both documents are found on the AHRQ QI website (www.qualityindicators.ahrq.gov) and provided in the excel spreadsheet provided with submission.

There are several steps involved in estimating the QI risk-adjustment models.

1. Construct candidate covariates
2. Select model covariates
3. Estimate the models
4. Evaluate the models

For the PQIs, potential risk-adjustment indicate whether the discharge record meets the technical specification for gender, age in 5-year groups and poverty category that are used as covariates in the risk-adjustment model.

Covariates are coded for each discharge record based on the data elements, data values, and logic described in the technical specifications and the appendices of the risk-adjustment coefficient tables. For a given covariate, if the discharge meets the technical specification for that covariate a value of "1" is assigned to the discharge level covariate data element. Otherwise a value of "0" is assigned to the discharge level covariate data element.

Risk Adjustment Coefficients as presented in the *PQI Parameter Estimates* document (pages 1-2) posted at:

http://www.qualityindicators.ahrq.gov/Downloads/Modules/PQI/V45/Parameter_Estimates_PQI_45.pdf

Table 1. Risk Adjustment Coefficients for PQI #1 Diabetes Short-Term Complications Admission Rate

PARAMETER	LABEL	DF	ESTIMATE	STANDARD ERROR	WALD CHI-SQUARE	PR > CHI-SQUARE
INTERCEPT		1	-8.0120	0.0416	37143.58	< 0.0001
SEX	Female	1	-0.1430	0.0519	7.65	0.0057
AGE5	Male, Age 18-24	1	0.9129	0.0425	460.50	< 0.0001
AGE6	Male, Age 25-29	1	0.8191	0.0431	361.07	< 0.0001
AGE7	Male, Age 30-34	1	0.8126	0.0432	353.71	< 0.0001
AGE8	Male, Age 35-39	1	0.7570	0.0433	305.53	< 0.0001
AGE9	Male, Age 40-44	1	0.8595	0.0431	398.02	< 0.0001
AGE10	Male, Age 45-49	1	0.8403	0.0430	381.81	< 0.0001
AGE11	Male, Age 50-54	1	0.7143	0.0432	273.06	< 0.0001
AGE12	Male, Age 55-59	1	0.5363	0.0438	149.78	< 0.0001
AGE13	Male, Age 60-64	1	0.3375	0.0448	56.84	< 0.0001
AGE14	Male, Age 65-69	1	0.2139	0.0465	21.18	< 0.0001
AGE15	Male, Age 70-74	1	0.1144	0.0488	5.48	0.0192
AGE16	Male, Age 75-79	1	0.1127	0.0511	4.86	0.0274
AGE17	Male, Age 80-84	1	0.0655	0.0548	1.43	0.2323
AGE5	Female, Age 18-24	1	0.3665	0.0534	47.17	< 0.0001
AGE6	Female, Age 25-29	1	0.0813	0.0545	2.23	0.1358
AGE7	Female, Age 30-34	1	-0.1260	0.0549	5.29	0.0215
AGE8	Female, Age 35-39	1	-0.0210	0.0549	0.16	0.6935
AGE9	Female, Age 40-44	1	-0.1800	0.0548	10.90	0.0010
AGE10	Female, Age 45-49	1	-0.1600	0.0546	8.67	0.0032
AGE11	Female, Age 50-54	1	-0.0630	0.0548	1.35	0.2454
AGE12	Female, Age 55-59	1	0.0553	0.0556	0.99	0.3192
AGE13	Female, Age 60-64	1	0.0926	0.0569	2.65	0.1036
AGE14	Female, Age 65-69	1	0.0852	0.0595	2.06	0.1517

(CONTINUED)

PARAMETER	LABEL	DF	ESTIMATE	STANDARD ERROR	WALD CHI-SQUARE	PR > CHI-SQUARE
AGE15	Female, Age 70-74	1	0.1170	0.0627	3.49	0.0619
AGE16	Female, Age 75-79	1	0.1120	0.0654	2.93	0.0870
AGE17	Female, Age 80-84	1	0.1198	0.0696	2.96	0.0852

c-statistic: Measures of association between the observed and predicted values were not calculated because the predicted probabilities are indistinguishable when they are classified into intervals of length 0.002.

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)

In general, for discrimination, we calculate the c-statistic by taking all possible pairs of cases consisting of one case that experienced the event of interest and one case that did not experience the event of interest. The c-statistic is the proportion of such pairs in which the case that experienced the event had a higher predicted probability of experiencing the event than the case that did not experience the event.

In general, for calibration, we assign each person to a decile based on the predicted rate from the risk-adjustment model, and calculate the average predicted rate and average observed rate per decile. A model that is well calibration will have observed values similar to predicted values across the predicted

value deciles. Although there are statistical tests of such “goodness of fit” the tests generally are not informative for datasets with large sample sizes.

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):

See 2b4.8 (Table 6)

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

See 2b4.8 (Table 6)

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

Table 6. Model Discrimination and Calibration

Predicted Rate Decile	Number of Persons per Decile	Predicted Rate	Observed Rate
1	23,686,174	0.000344	0.000359
2	23,684,650	0.000403	0.000423
3	23,709,985	0.000490	0.000504
4	23,660,585	0.000559	0.000600
5	23,702,124	0.000567	0.000630
6	23,669,467	0.000625	0.000641
7	23,684,401	0.000707	0.000754
8	23,690,076	0.000752	0.000761
9	23,685,783	0.000791	0.000830
10	23,680,145	0.000955	0.000977
C-statistic	Not Calculated		

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.

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)

The model has poor discrimination based on a norm c-statistic of 0.80 but moderate calibration. Measures of association between the observed and predicted values were not calculated because the predicted probabilities are indistinguishable when they are classified into intervals of length 0.002.

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

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 county (area) parameterized using the Gamma distribution. We then calculate the probability that the county (area) is better or worse than the reference population rate at a 95 percent probability overall and by area size decile. The analysis is with the computed 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 7. Performance Categories by County (Area) Size Decile

Size Decile	Number of Counties (Areas)	Ave. Number of persons per County (Area) in Decile	Proportion Better	Proportion Worse
1	312	2,278.3	0.22115	0.10256
2	311	5,657.8	0.40836	0.17685
3	311	8,817.1	0.36334	0.27331
4	311	12,641.0	0.32476	0.27974
5	311	17,289.0	0.36656	0.31833
6	312	23,989.3	0.28846	0.39744
7	311	33,768.0	0.32797	0.39228
8	311	53,199.9	0.31833	0.45016
9	311	103,760.9	0.35691	0.45016
10	311	500,100.9	0.41158	0.45659
	3,112	76,109.7	0.33869	0.32969
Patient weighted			0.42409	0.43578

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?)

Counties with large population are more likely to be identified as better as or worse than the reference population rate due to the lower uncertainty in the performance score.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, 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)

Not applicable

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

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