

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

Measure Number (if previously endorsed): Click here to enter NQF number

Measure Title: Click here to enter measure title

Date of Submission: Click here to enter a date

Type of Measure:

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

Data come from two sources:

- OptumInsight (formerly known as Integrated Healthcare Information Services, Inc. (IHCS)) research database, used to develop and test methodology and measurement approaches.
- Health plan reported data as part of the HEDIS measurement program.

To guide the development of the RRU measures (2004–2012), NCQA convened an expert advisory panel, the Efficiency Measurement Advisory Panel (EMAP) to discuss different methodological issues related to RRU measurement and develop an approach to measure relative resource use.

Using a large managed care database from OptumInsight (formerly known as Integrated Healthcare Information Services, Inc. (IHCS)), NCQA performed extensive research on the different methodologic and measurement approaches issues proposed by the EMAP.

In addition, NCQA annually conducts an analysis on the data submitted for the HEDIS RRU measures, including an examination of the reliability and validity of the current year data compared to all previous years' data. The intent of this annual report is to ensure the continued reliability and consistency of the data used to calculate the RRU results. The primary data for the most recent analyses are the HEDIS 2012 reports of relative resource use (RRU) by commercial, Medicaid, and Medicare plans. These results are reviewed by the Efficiency Measurement Advisory Panel (EMAP) and results are approved by the

Committee on Performance Measurement. A standard set of questions are asked to ensure the validity and repeatability of the RRU results that are publically reported.

1.3. What are the dates of the data used in testing? 2004-2012

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 checked="" type="checkbox"/> health plan	<input checked="" 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)

1) Optuminsight RRU Research Database = 25 million unique individuals, over 44 health plans and other contributors.

2) The RRU data used for the annual reliability and stability analyses are drawn from all HEDIS health plan submissions for the 2012 calendar year (commercial =359 plans, Medicaid =86 plans, and Medicare =219 plans).

The IHCIS-Managed Care Benchmark Database includes medical and pharmacy claims and enrollment for more than 25 million unique individuals, 30 health plans and other contributors. The database population was comprised of primarily non-elderly, commercially enrolled individuals. All data were standardized and evaluated for completeness and consistency. Costs were based on a standard pricing methodology applied across all contributors and time periods (using Relative Value Units (RVUs) and other methodologies). For the analysis described here, a subset of the Benchmark Database population was selected. In particular, the study population met the following criteria:

at least 6 months of enrollment in the year (2003) used to identify patients and measure costs and utilization.

selected from a number of different populations (health plans) that met sufficient product and geographic variation (given available data).

In the end 1 Medicare Risk, 1 Medicaid and 12 commercial populations were selected for the initial study meeting the above selection criteria.

The primary data for most recent annual analysis are the 2012 HEDIS national submissions of relative resource use (RRU) by commercial (359), Medicaid (86), and Medicare (219) plans.

Formatted: Font: 11 pt

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)

At the time of initial RRU development testing, the total population meeting the OptumInsight (IHCIS Managed Care Benchmark Database) criteria exceeded 7.5 million individuals. The population included a mix of HMO, PPO and POS products and included Blue Cross Blue Shield and regional plans of different sizes from across the U.S. This database has been regularly updated and the more recent analysis were conducted on a total population of 25 unique individuals across 44 health plans

Formatted: Font: 11 pt

~~At the time of testing, the total population meeting the IHCIS Managed Care Benchmark Database criteria exceeded 7.5 million individuals. The population included a mix of HMO, PPO and POS products and included Blue Cross Blue Shield and regional plans of different sizes from across the U.S. Annual analysis data is drawn from all HEDIS health plan submissions for the commercial, Medicare and Medicaid product lines. Most recent data (2012)~~
The 2012 HEDIS RRU data reports relative resource use for approximately 1,913,266 male and 2,210,579 female patients between the ages of 18 and 85 with cardiovascular conditions across these three product lines.

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.

None

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)

NCQA's Relative Resource Use (RRU) measure for People with Cardiovascular Conditions has undergone multiple levels of reliability and validity testing to ensure that the measure results represent meaningful information on the resources used by a health plan to manage its members with asthma. The testing can be broken down into three major types: 1) Development Field Test, 2) Implementation Feasibility testing (including reliability of selected data elements), and 3) Annual Analysis of RRU data. Each of these testing types will be described below, in section 2a2.3 and 2a2.4 and in further detail in attachments SA Reliability Validity+Testing.pdf, and SA Standardized Price Implemetnation.pdf.

Formatted: Font: 11 pt

Formatted: Font: 11 pt

Resource Use: In 2003, NCQA began to investigate several strategies to measure cost and resource use for patients with specified conditions. The goal was to develop a measurement strategy that would accurately and reliably capture the resources used for patient populations by service category. The proprietary nature of prices and discounts negotiated between health plans and providers led us to a standard costing methodology. Costs were aggregated using service counts and RVU per service in order to convert RVU to a relative dollar amount. Pricing levels reflect total allowed payments, inclusive of health plan liability and patient cost-sharing and reported by per patient per month (PMPM). (For details see SA Reliability Validity+Testing.pdf attachment.)

Data Element Reliability: The implementation and feasibility study is illustrative of how NCQA examines the consistency of service claims for Relative Resource Use measures using a cross sample of health plan member data. Most recently NCQA looked to add Diagnostic Laboratory and Imaging service categories to the RRU measurement set and needed to confirm that these services were being coded with adequate consistency and reliably for us to generate a reliable standard price assignment for each individual coded service. IN a sample of health plan members' data from 40 health plans, we cross referenced up to 2 records per day per member for revenue codes, CPT-global codes and CPT codes with either TC or 26 modifiers present. Consistency of administrative claims were assessed for lab and imaging by looking at each of the following scenarios for each member in the sample:

- Single and multiple claim record scenarios for coding and place of service (POS).
- Single and multiple claim record scenarios with respect to radiology service records.
- The distribution of single images and multiple views to determine the consistency of pricing.
- distribution of scenarios for one vs. multiple rows of revenue codes for imaging
- The variation in Imaging Cost per Service, by Revenue and by Coding Scenarios.
- The usage of modifiers (26 or TC) for variation across plans and/or correlation with corresponding RRU results.

Full results of this investigation can be found in the attachment
SA Standardized Price Implementation.pdf.

Annual RRU Analysis: Every year since the HEDIS RRU measures were approved for public reporting in 2009, NCQA has analyzed the data submitted to evaluate the continued reliability and consistency of the data used to calculate the RRU results. The primary sources of data for the most recent analyses are the 2012 submissions of HEDIS RRU measure results by product line (Commercial, Medicare and Medicaid), and are comprised of cumulative plan observations across all data dimensions (e.g., product line, reporting type). The relationships are examined and cross referenced at each component level for positive and negative correlations (Absolute value of Spearman correlation coefficient). NCQA utilizes these analyses to examine the distribution of submitted plan data and the subsequent observed-to-expected ratios. These results are reviewed by the Efficiency Measurement Advisory Panel (EMAP) and subsequently submitted for review and approval by the Committee on Performance Measurement. A standard set of questions are asked to ensure the validity and repeatability of the RRU results that are publically reported, and measures are not collected until approved by NCQA's Board of Directors.

NCQA annually conducts an analysis on the data submitted for the HEDIS RRU measures, including an examination of the reliability and validity of the current year data compared to all previous years' data. The intent of this annual report is to ensure the continued reliability and consistency of the data used to

Formatted: Font: 10 pt

Formatted: Font: 11 pt

Formatted: Font: 12 pt

Formatted: Font: 11 pt

calculate the RRU results. The primary sources of data for these analyses are the 2012 submissions of HEDIS RRU measure results by product line (Commercial, Medicare and Medicaid), which are compared with submissions from prior years. Reports for corresponding HEDIS Effectiveness of Care (EOC) quality measures from 2012 and years prior were also used to examine co-variation in performance. Health plans estimate standardized (in dollars) resource use by following HEDIS specifications, including applying the NCQA standardized prices for each unit of health service included in each measure. NCQA estimates the “expected” standardized utilization by calculating the average utilization within each clinical reporting category for each health plan submitting data.

NOTE: NCQA will be pleased to share with the Steering Committee the “RRU Annual Report” and the “SA Reliability Validity Testing” however due to page restrictions we cannot attach to this testing form or the submission form.

The analyses have been structured to provide comprehensive univariate (descriptive) information and selected correlation results. The analyses use cumulative plan observations across all data dimensions (e.g., product line, reporting type). NCQA utilized these analyses to examine the distribution of submitted plan data and the subsequent observed to expected ratios. We also examined the number of plans that were successfully able to report RRU data and compared to previous years’ performance. Plans with estimated O/E results less than 1/3 or greater than 3.000 or incomplete data for corresponding HEDIS EOC measures were defined as unable to successfully report RRU data.

NCQA was able to set specific objectives for the 2012 RRU analysis to examine the continued reliability and validity of the RRU HEDIS data supporting the measures:

- Are a sufficient number of plans reporting RRU data?
- Did notice of public reporting of RRU results in 2012 result in a change in the number of makeup of plans that reported RRU in 2012?
- Has the range in RRU results remained stable over time?
- Did the number of plans identified as “outliers” change in 2012?
- Are plans’ observed to expected results for the RRU measures stable over time?
 - Stability over time indicates that spurious observations and results are not common and that estimates of resource use are stable over time.
 - Resource use for individual plans should not change appreciably.
- Is there a relationship between plans’ O/E results and quality results? There are few significant correlations between risk-adjusted resource use and the quality composite, all of which are weak at best.

Additional reliability testing was performed on the data that are used to report relative resource use for people with cardiovascular conditions. For example, the most recent addition of diagnostic laboratory and imaging services required an analysis of the coding reliability for each of these services across different plan environments to ensure that both the standard pricing methodology and the reporting would accurately reflect the level of resources used to manage patients. Claims data from 40 health plans were reviewed and analyzed for a variety of revenue and procedure codes with technical (TC) and professional (26) modifier combinations. Each of the service records examined were filed on the same date of service for the same member in order to determine the amount of “noise” due to different

methods of data capture. Up to 2 records per day per member were cross-referenced for revenue codes, CPT global codes and CPT codes with either TC or 26 modifiers. The results of this analysis, supported by data, indicating the following:

- Instances with both a technical and professional claim for the same CPT code and appropriate modifier can be priced reliably regardless of the place of service or provider.
- Professional CPT/modifier-coded services are frequently used; however, the technical component is coded less frequently due to place of service issues. Subsequently, the professional component can be priced with more ease than the technical component.
- For many lab-related CPT codes, modifiers are used sparsely given that professional services are not expected to be utilized

Correlations between components of the RRU measurement set and their corresponding HEDIS EOC quality indicators generally provide an indication of value in terms of the resource use quality relationship. For these analyses, the relationships were defined as moderate to strong positive correlation (Absolute value of Spearman correlation coefficient > 0.30 with a p -value < 0.01) or moderate to strong negative correlation (Spearman correlation coefficient < -0.30 with a p -value < 0.01). NCQA assessed the association between the RRU O/E ratios and the quality composite, as well as the indexed RRU O/E ratios and the indexed quality composite. The relationships between cost and quality were determined to be the same regardless of whether indexed or non-indexed values were used.

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)

Results from the 2012 analyses indicate that plan performance generally remained stable across the 2011-2012 reporting period, although eligible population size continues to be a leading factor in preventing public reporting of several plans' RRU results. In 2012 a total of 728 HMO and 328 PPO plans submitted HEDIS data to NCQA. With respect to RRU-specific HEDIS data, a total of 428 HMO and 244 PPO plans submitted data which is approximately 82% of commercial HMOs and 94% of commercial PPOs reporting any HEDIS data, 48% of Medicare HMOs and 48% of Medicare PPOs, and 49% of Medicaid HMOs. In general, the trend in number of plans reporting RRU data is directly related to the number of plans reporting any HEDIS data. Additionally the increase in the number of Medicare plans reporting RRU data from 2011 to 2012 is consistent with the increase in the total number of plans seen reporting any HEDIS data.

Additionally, correlation analyses (cost component-performance; cost component-quality; and component-component) demonstrated the increased precision of the updated risk adjustment model.

Below are the key findings from this year's report.

- HMO plans continue to have higher submission counts for RRU-specific data than PPO plans.
- The majority of plans' observed-to-expected (O/E) results for both the *Total Pharmacy* and *Total Medical* categories stayed within or moved no more than one quartile regardless of product line, reporting type or disease condition.
- Returning plans are still performing better than new plans in terms of meeting criteria required for public reporting (outlier distribution and data completeness) of RRU results.

The majority of observed new correlations considered 'moderate to strong positive' included the *Procedures and Surgery (aggregated)* component, and the majority of observed new correlations considered 'moderate to strong negative' included the *Total ED Discharges* component. Other correlations, both moderate to strong positive and moderate to strong negative, remained unchanged in terms of occurrence and directionality or subsided altogether from 2011 to 2012 across all measures.

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?)

In addition to the number of plans submitting HEDIS and/or RRU-specific data, it is useful to examine how many plans meet public reporting status based on pre-defined criteria including: O/E ratio outlier distribution (for a specific service category or overall), data completeness for corresponding HEDIS EOC measure submissions and minimum eligible population size requirements. Plans are eliminated from public reporting in the *Total Medical* category if 1) any of the O/E ratios for components of *Total Medical* or *Total Pharmacy* is less than 1/3 or greater than 3.000; 2) if their eligible population size for a measure is less than 250, or 3) if the plan is missing required quality data in their HEDIS EOC measure submission file. Due to the increase in plans reporting RRU for the first time and the effect they might have on the reported results (assuming a steep learning curve for new plans), additional analyses are conducted based on their assigned reporting status ("New" submissions against "Returning" RRU plans). Status is assigned as "New" by determining whether the health plan has any RRU data associated

Eligible population size is the primary reason that plans do not qualify for public reporting status, regardless of product line, reporting type or clinical condition. However there have been instances of improvement in meeting this criterion, especially in terms of reporting for the CV Conditions measure. For the CV Conditions measure, the percentage of plans eliminated due to eligible population size decreased from 67.7% in 2011 to 44.8% in 2012 for Medicare PPO and from 43.5% in 2011 to 21.7 in 2012 for Medicare HMO; and from 70.4% in 2011 to 42.9% in 2012 for Medicaid HMO.

In terms of O/E ratio outlier distribution for the cardiovascular conditions RRU measure, no commercial plans were eliminated from *Total Medical* O/E results falling outside the pre-defined outlier range. In the Medicare and Medicaid reporting lines, approximately 2% of plans were found to have *Total Medical* O/E results below the 0.333 outlier threshold.

An indicator of plan stability over time is quartile movement of O/E ratios (for specific and overall service categories), with significant shifts having implications about plan performance in terms of resource use. For comparative purposes, plans that move less than one quartile are considered stable, with the magnitude of absolute change being more relevant as opposed to the direction of change (up or down). Since the benchmarks are calculated every year based on total RRU submissions, a single plan's O/E ratio can move about the mean without having any significant change in their observed data from one year to the next. For each RRU measure NCQA calculated the percentile distribution for each cost component (by product line, reporting type and year) and then determined the quartile into which each plan's O/E ratio fell for a given cost component. NCQA then examined if the plan's 2012 O/E ratio remained within at least one quartile of the previous years. This analysis should reveal any erratic shifts in plan performance across years, which could indicate instances of data capture, data error or

insufficient case-mix adjustment. Overall the majority of plans' O/E ratios for *Total Pharmacy* and *Total Medical* stayed within or moved no more than one quartile between successive years, regardless of product line, reporting type or clinical condition.

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)

Method of Assessing Face Validity: NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle, which is outlined below. Our Measurement Advisory Panels (e.g., the Efficiency Measurement Advisory Panel and our Risk Adjustment Advisory Panel) and our Technical Panels (e.g., Pharmacy Panel, Coding Panel, Lab Panel) operate on a consensus basis to encourage ongoing work to both develop new measures and improve them over time. Our Committee on Performance Measurement (CPM) is a committee of NCQA's Board of Directors and has been in continuous service for 20 years. The CPM votes to approve all measures included in NCQA programs including HEDIS (Health Plan, ACO, and Physician), as well as measures used in Physician Recognition Programs. A quorum (50% of the members + 1) must be present during discussion to vote for a measure. A majority must vote in favor of a measure to be approved. A tie vote does not approve the measure. NCQA does not release specific voting results of the Board or its respective Committees.

Formatted: Font: 11 pt

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format and vetted by NCQA's Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary. To guide the development of the RRU measures, NCQA convened an expert advisory panel, the Efficiency Measurement Advisory Panel (EMAP) (See Section Ad.1 of submission form for a list of the EMAP and CPM members) to discuss different methodological issues related to RRU measurement and develop an approach to measure relative resource use.

Formatted: Font: 11 pt

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing

health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publically reported and may be used for scoring in accreditation. Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed.

Method of Assessing Empirical Validity¹: For the developmental phase of the RRU measures (2003-2005), we wished to know what the typical total expenditures was for patients with different chronic conditions. To do this, cost and utilization experience were measured for the same 12 months used to identify patients. All inpatient facility, outpatient facility, professional, ancillary and pharmacy claims for the disease-identified members were selected. The selected service categories included inpatient facility, pharmacy, evaluation and management (including consults), procedures (including outpatient facility and ambulatory surgical center services), laboratory, and imaging services. The cost measure used in the analysis was based on a standard costing methodology and priced at calendar year (CY) 2003 levels. For the purposes of the developmental field test, pricing levels reflect total allowed payments, inclusive of health plan liability and patient cost-sharing. Costs were reported on a cost per patient per month (PMPM) basis. Since a standard costing methodology was employed for the field test study data, the costs reported can be considered "weighted utilization," i.e., they were computed using service

Formatted: Font: 11 pt

Formatted: Font: 11 pt

¹ More detailed results can be found in section 2b2.3. with additional results in Tables 5, 6 and 7 of the Attachment [SA Reliability Validity+Testing.pdf](#)

counts and RVUs per service and a dollar factor to convert RVUs to dollars. These RVUs represent units of standard priced dollars, in relative terms.

This measurement required a population-based risk assessment approach that could capture the overall patient morbidity, including conditions related to the clinical category being studied as well as all conditions observed for the patient. Morbidity categories include groups of patients with similar levels of health risk. Initially, two different approaches were used to assign patients to morbidity categories for the analysis. The first method employed Episode Risk Groups (ERGs). The second approach to morbidity adjustment for measuring the relative resource utilization for total service employed an age-sex model. For specific information on the development of the current risk adjustment approach please refer to attachment SA RA Feasibility Resource Burden FTR

For the developmental phase of the RRU measures, cost and utilization experience were measured for the same 12 months used to identify patients. All inpatient facility, outpatient facility, professional, ancillary and pharmacy claims for the disease-identified members were selected. Measures of cost and utilization were produced for all services and some selected service categories that may serve as a proxy for all services. The selected service categories included inpatient facility, pharmacy, evaluation and management (including consults), procedures (including outpatient facility and ambulatory surgical center services), laboratory, and imaging services. The cost measure used in the analysis was based on a standard costing methodology and priced at calendar year (CY) 2003 levels. For the purposes of the developmental field test, pricing levels reflect total allowed payments, inclusive of health plan liability and patient cost sharing. Costs were reported by a cost per patient per month (PMPM) measure. Since a standard costing methodology was employed for the field test study data, the costs reported can be considered "weighted utilization," i.e., they were computed using service counts and RVUs per service and a dollar factor to convert RVUs to dollars. These RVUs represent units of standard priced dollars, in relative terms.

Early in measure development (2004-2005), two different approaches were tested to identify disease-related costs. The first approach employed a widely-used tool, ETGs, which uses an episode of care approach to assign medical and pharmacy services to conditions and diseases. Episodes are created based on a series of rules and the diagnoses and procedures found on medical claims, including drug treatments listed on 14 pharmacy claims. For this field study the ETG grouper software was applied to 12 months of medical and pharmacy claims used for each patient. The result was an output file that includes the ETG assigned to each service, along with other information, which were then mapped to each of the major clinical groupings. Where patients were identified for a clinical grouping within a larger major clinical category (e.g., cardiovascular or asthma/COPD), all of the disease-related costs within that category were assigned as disease-related for that clinical grouping for that patient. The same approach was used for asthma/COPD, where a patient identified ultimately as a COPD patient received the disease-related costs for both asthma and COPD. Since ETGs assign each service uniquely to a single episode of care, services could not be disease-related to multiple major clinical categories.

The second approach to assigning disease-related costs employed Disease Identification (DID) approach. A medical service was determined to be disease-related if any of the diagnosis (using the first 3

diagnostic positions) or procedure codes on the service corresponded to one or more of the diagnosis or procedure codes used to identify the clinical categories. Disease-related pharmacy services were identified based on the NDC code on the pharmacy claim and were mapped to the highest-level therapeutic categorization developed for each major clinical category. Since a single service could have multiple diagnosis codes (some of which could be assigned to a different clinical category), using the DID approach allows a service to be used as disease-related for multiple conditions.

The disease-related methodologies were used to assign services and costs to each clinical category. An important objective of the study was also to measure total service costs for patients in each clinical category, including those related to the disease and other services. This measurement required a population-based risk assessment approach that could capture the overall patient morbidity, including conditions related to the clinical category being studied as well as all conditions observed for the patient. Morbidity categories include groups of patients with similar levels of health risk. Two different approaches were used to assign patients to morbidity categories for the analysis. The first method employed Episode Risk Groups (ERGs). A risk score of 0.50 indicates a health risk approximately half of that of the average member in an index population, a score of 1.00 means the patient's relative risk is equal to the average member, and 1.50 indicates a fifty percent greater risk. Retrospective (concurrent) values of health risk were used for the analysis. Eight ERG morbidity categories were created for use in the study:

risk score less than 1.00 5. risk score 8.00 to less than 12.00

risk score 1.00 to less than 2.00 6. risk score 12.00 to less than 15.00

risk score 2.00 to less than 4.00 7. risk score 15.00 to less than 20.00

risk score 4.00 to less than 8.00 8. risk score 20.00 or higher

Using their risk score a patient was assigned to the appropriate ERG morbidity category. The ranges used for these categories were based on the observed distribution of risk for study patients and the desire to create a limited number of categories to support sufficient sample size within each grouping and also to limit reporting burden. The second approach to morbidity adjustment for measuring the relative resource utilization for total service employed an age-sex model. Based on an analysis of the distribution of study patients and their costs, the following age-sex categories were employed, where "All" indicates both genders for the same age range:

All, 00-17 years

Females, 18-44 years

Males, 18-44 years

All, 45-54 years

All, 55-64 years

All, 65-74 years

All, 75+ years

~~In summary, ERGs and the age-sex model were used as the basis for creating morbidity categories to support total service measurement. Further, given the stratification of patients into the 18 clinical categories previously described, the final population-based risk assessment methodology was an ERG-based Morbidity Adjustment — using ERGs within clinical categories, including with and without co-morbidity alongside an “Age-Sex” and Clinical Category-based Morbidity Adjustment — using age-sex groupings, within clinical categories, including with and without co-morbidity.~~

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

The investigations provided insights into the conceptual and methodological issues in measuring relative utilization at a health plan level. Using a large research database, the study addressed a number of questions related to assessing resource utilization at the health plan and population levels. The following questions were assessed during the initial validity testing of the RRU approach:

Question 1: What is the typical total expenditures for patients with different conditions? Do patients with the same condition and co-morbidity have different costs? How do the estimates vary across populations?

Interpretation (See Table 5, page 35 of attached file “SA_Reliability Validity Testing”):

- Patient costs were highest for AMI and CHF and lowest, on average, for asthma patients.
- As expected, costs for members with a condition and a qualified co-morbidity were higher than for patients with the same condition without co-morbidity.
- In general (with a few exceptions), the average costs for a clinical grouping were similar across plans.

Question 2: What is the typical total expenditures for patients with different conditions, by service category? What is the most important service category financially? How do the estimates vary across clinical categories?

Interpretation (See Table 6, page 36 of the file “SA_Reliability Validity Testing”):

- As expected, variation in patient costs across clinical categories was observed. Further, differences in the relative importance of categories by clinical grouping were also evident.
- Inpatient and pharmacy services comprise the largest individual service category percentages. Inpatient services were most important for cardiovascular conditions.
- The “Other” category (denoting services that may be more difficult to quantify and measure) comprises 10-15 percent of total service costs – a consistent percentage across clinical groupings.

Question 3: What is the magnitude of disease-related costs for each clinical grouping? How do these amounts vary by service category?

Interpretation (See Tables 7&8, pp. 38-40 of the file “SA_Reliability Validity Testing”):

- Disease-related costs represent a significant portion of total service costs for some conditions
 - in particular the cardiovascular conditions (approx 50-80 percent). These percentages vary by service category.
- Disease-related costs represent a lesser portion of total service costs for some conditions, e.g., asthma, COPD, arthritis and LBP.

- For many conditions, the magnitude of the disease-related costs was comparable whether using the ETG or DID approach – the exceptions were asthma, COPD and diabetes, with comorbidity, where the DID amounts were higher (for total services and other service categories). In general, findings were comparable between the two approaches.

Findings on Relative Resource Utilization – Variation by Type of Service:

- For a given health plan and clinical category, measures of relative resource utilization were generally similar across different types of service, with only some modest variations. The consistency was greatest for those services comprising a larger portion of overall costs measured (e.g., inpatient and pharmacy) in addition to showing the variation in findings across type of service categories.
- For a given health plan and clinical category, measures of relative resource utilization were generally similar using the “selected” group of services (inpatient, pharmacy, E&M and procedures) versus all types of service. In general, where differences were observed, relative resource utilization for diagnostic services (radiology, laboratory, and other diagnostic testing) were the primary factor.

The study explored the potential for the use of a subset of services as a proxy for measuring resource use for all services (see Table 7 pp. 38 of the file “SA_Reliability_Veracity Testing”). In this way, services that can be reliably measured could be the focus of initial measurement and also present a reasonable burden on health plans in collecting this information. The study found measures of relative resource utilization were generally similar using “selected” services (inpatient, pharmacy, evaluation and management, and procedures, including ASC costs) versus measurement using all services.

Findings on Relative Resource Utilization – Variation across Clinical Category:

For a given population, measures of relative resource utilization were generally similar across the major clinical categories, i.e., similar findings were observed for the same population for cardiovascular disease, diabetes, depression, asthma/COPD, and arthritis/LBP. This was particularly true for total service costs. For disease-related costs somewhat greater variation was observed across conditions for the same population.

Findings on Relative Resource Utilization – Variation across the Four Methods

For a given population and clinical category, measures of resource utilization were generally similar across the four different approaches to measurement described above, with only some modest variations.

Summary Interpretation:

- A typical standard error for measuring total service relative resource utilization was observed to be approximately 0.025 at samples of 2,000 patients or more. For example, for a condition with a typical prevalence of 1 percent of enrolled members, a health plan of 250,000 members would yield a patient sample of 2,500. Based on the above standard error, the expected 95 percent confidence interval around the estimated resource utilization index would be approximately +/- 0.05, where 0.05 equals twice 0.025 (a 95 percent confidence interval is approximately 2 standard errors).

- In general, the standard errors were relatively higher for measures of disease-related services versus total services.

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?)

The Measuring Health Plan Relative Resource Utilization study (2005) produced a number of key findings related to resource measurement; however, we are still challenged by the value of these metrics and their meaning to purchasers. The study conclusively determined that:

- Health plans can be meaningfully measured and compared with respect to the relative resource consumption of their networks for select resource categories.
- Methodologically defensible non-proprietary methods can be identified for severity and case adjustment. These methods can serve as the basis for the development of practical algorithms to support measurement of resource utilization at the health plan level – involving a reasonable burden on health plans in measurement and also avoiding the need for requiring their use of a proprietary tool.
- A significant obstacle in sharing cost information at the health plan level is the proprietary nature of the fee schedules and contracts that describe their pricing of services. This study employed standard pricing methods that removed unit price variation as a factor in resource measurement.
- Relative resource consumption seems to vary meaningfully between health plans. More specific findings related to these measures provided insights related to the services, conditions and methods used for study:
- Services – for a given health plan and clinical category, measures of relative resource utilization were generally similar across different types of service, with only some modest variations. The consistency was greatest for those services comprising a larger portion of overall costs measured (e.g., inpatient and pharmacy).
- Study Conditions – for a given health plan, measures of relative resource utilization were generally similar across the study conditions – i.e., similar findings were observed for the same population for cardiovascular disease, diabetes, depression, asthma/COPD, arthritis and LBP.
- Methods – four different approaches were used by the study to measure relative resource use – varying by the risk adjustment methodology employed and the focus on total service versus disease-related costs. For a given population and clinical category, measures of resource utilization were generally similar across the four different approaches to measurement described above, with only some modest variations.
- The study explored the potential for the use of a subset of services as a proxy for measuring resource use for all services. In this way, services that can be reliably measured could be the focus of initial measurement and also present a reasonable burden on health plans in collecting this information. The study found measures of relative resource utilization were generally similar using “selected” services costs) versus measurement using all services.

The relationship between population size and variation in measures of relative resource utilization – i.e., what is a sufficient sample size to produce consistently valid numerators and denominators and how large of a health plan is required to achieve these thresholds – was explored.

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)

Measure specifications require that members of plans in all three product lines who had evidence of other dominant medical conditions, such as active cancer, specific organ transplants (non-renal), or HIV/AIDS, are required to be excluded from RRU measurement. Patient age criteria are also used to exclude individuals, specifically: patients less than 18 years of age or greater than 75 years of age are excluded from cardiovascular conditions.

ESRD and renal transplants are particularly relevant to management of patients affected by diabetes and cardiovascular (CV) conditions. Cardiovascular disease disproportionately affects patients with CKD, with 26% of patients with CKD having a co-morbidity of cerebrovascular accidents and transient ischemic attacks; 12.5% with acute myocardial infarction; and 43.6% with congestive heart failure in 2010.

Methodology: OptumInsight evaluated the prevalence and costs associated with ESRD and renal transplants for the RRU eligible population and specific cohorts of patients. The investigation involved the following steps:

1. Segmentation of the RRU research database by disease and risk cohorts
 - a. **Cohort 1:** All members for condition, including members who will be excluded
 - b. **Cohort 2:** All members after all exclusions applied
 - i. Patients with dominant conditions of active cancer, transplant status (renal), ESRD and HIV/AIDS **not included**
 - c. **Cohort 3:** All members after exclusions for ESRD and transplant status applied
 - i. Patients with dominant conditions of active cancer and HIV/AIDS **not included**
 - d. **Cohort 4:** Members with ESRD and transplant status (renal)
 - i. Patients with dominant conditions of ESRD and transplant status (renal) **included**
2. Summary of costs (per member per month) for cohorts based on exclusions (including for ESRD and renal transplant where 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)

Overall ESRD and transplant status (renal) comprised small percentages of the CV conditions patient populations (1.5 and 1.0 respectively). However both were major contributors to costs incurred regardless of the primary condition or service category in consideration. Results illustrate that although the prevalence of the CV conditions with a co-morbid condition of ESRD and transplant status (renal) is small, these services contribute significantly to costs associated with medical care (Table 1).

Table 1: Relative Cost (PMPM) Ratio of Cohort 4 to Cohort 3, CV Conditions

	Evaluation & Management	Procedure & Surgery
--	-------------------------	---------------------

Age	Gender	# Members	Medical	Pharmacy	Inpat	Outpat	Inpat	Outpat	Total Pharmacy	Inpat Facility	Imaging	Lab	Total Medical	Total
18-44	F	56	672	396	4.5	1.3	4.5	3.1	2.1	4.9	1.3	2.9	3.9	3.6
18-44	M	81	971	540	4.8	1.8	4.0	5.9	2.8	4.7	2.3	4.4	4.3	4.0
45-54	F	241	2,892	1,680	4.5	1.5	3.8	2.9	2.6	4.4	1.6	3.5	3.6	3.3
45-54	M	400	4,797	2,914	5.7	1.9	3.8	4.4	2.2	5.0	2.5	4.4	4.4	3.7
55-64	F	566	6,791	3,851	5.3	1.5	3.8	3.2	2.0	4.9	1.6	3.0	3.8	3.3
55-64	M	1,234	14,805	8,901	6.3	1.8	3.8	3.5	2.1	5.7	2.0	3.5	4.5	3.7
Total		2,578	30,928	18,282	5.79	1.73	3.83	3.57	2.17	5.27	1.97	3.56	4.27	3.65

Cohort 3: All members after exclusions for ESRD and transplant status applied (Patients with dominant conditions of active cancer and HIV/AIDS not included)

Cohort 4: Members with ESRD and transplant status (renal) (Patients with dominant conditions of ESRD and transplant status (renal) included)

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*)

As outlined in Section 2b3.2, although the prevalence of comorbidities of ESRD in the CV conditions population is small, the services rendered to these patients could be significant in terms of the overall resource use provided to the population being measured. NCQA's Relative Resource Use measures standard pricing methodology includes a "safety valve" or cost cap for any member that has extraordinarily high utilization for any particular measurement period. It was determined through testing the effect of these exclusions that the proportion of patients with these comorbidities that are included in the Total Medical do not disproportionately affect the overall performance. These findings resulted in ESRD and Renal Transplants being removed as mandatory exclusions from the CV conditions RRU measure.

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 184 risk factors
- ☒ Stratification by 13 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.

RCA results are risk adjusted using the HCC-RRU methodology described in section 2b4.3

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)

The current risk model utilized by NCQA is based on components of the CMS-HCC risk adjustment methodology and accounts for age, gender, and HCC-RRU risk classifications that predict cost variability. For each condition, members are assigned to a clinical cohort category that provides a more specific classification of the condition based on diagnosis codes that are identified in claims for the member in the prior year. A member's age, gender, and HCC category determines their risk score (cohort). NCQA then calculates the average per-member per-month (PMPM) cost for each cohort then weights that cost by the total member months within each cohort. Each plan will have its own weight for each cohort since case-mix varies across plans. These weighted cohort PMPMs are then summed across all cohorts to arrive at a PMPM that would be expected if the "average" plan had the same case-mix as the plan in question. The ratio of the observed-to-expected PMPM utilization indicates the degree to which a plan deviates from expected performance. This is known as indirect standardization.

Health plans submit the member month and summarized standardized cost separately for each member cohort, and NCQA calculates expected per member per month (PMPM) results. Thus, each health plan's RRU results are adjusted based on its mix of members.

Selection of a risk approach for RRU measures involved comparing the precision of member level risk assessments using individual adjusted R-squares and estimation of the absolute difference in health plan O/E resource use results using three different risk adjustment models. Initially, the following four approaches were considered for the comparative analysis:

- Model 1: Initial (NCQA Age-Sex-Disease)
 - HEDIS RRU 2007/2008 risk adjustment approach
- Model 2
 - newly created markers of risk, including interactions
 - NCQA estimates risk weights
 - NCQA provides approach for health plan to summarize and assign risk score and level to each patient – using markers and weights
- Model 3: Use the age and sex and clinical factors available for the current approach, plus selected other clinical factors, model in a multivariate framework
 - $\text{Risk} = \sum a_s * \text{AgeSex}_s + \sum c_c * \text{ClinicalFactors}_c + \sum n_n * \text{NewClinicalFactors}_n$
 - The a_s , c_c and n_n parameters are the risk weights assigned to each age-sex or clinical factor – risk score for a patient is the sum of these weights for all factors observed for them
 - Also, potential for including interactions
- Model 4: HCC-RRU approach
 - Uses HCC-CMS clinical logic markers and algorithms – download code from CMS
 - NCQA estimates risk weights – includes clinical variables only and adds RRU Age-Sex factors
 - NCQA provides approach for health plan to summarize and Assign risk score and level to each patient – using markers and weights.

Stratification of RRU Results

NCQA collects resource measures at the plan level and summarizes across reporting cohorts along the following dimensions:

- a) Product line (3 levels): commercial, Medicaid, and Medicare;

- b) Reporting type (2 levels): HMO and PPO;
- c) Area level (2 levels): national and regional;
- d) Resource use or utilization (11 levels): inpatient facility, procedure and surgery (inpatient and outpatient), evaluation and management (inpatient and outpatient), laboratory services, imaging services, ambulatory pharmacy, inpatient discharges, emergency department discharges.

Stratification of RRU results to control for individual confounding variables is not performed since age, gender and risk variables (comorbidity and disease interactions) that affect healthcare costs are adjusted for in the RRU-HCC risk adjustment process. These include age and gender along with one of the 13 assigned HCC-RRU risk categories (e.g. male 18-44 HCC-RRU 1; male 18-44 HCC-RRU 2; male 18-44 HCC-RRU 3; etc...). However, in order to assist organizations in identifying opportunities for improvement, NCQA reports RRU results using the HCC-RRU cohorts as reporting strata. Reporting the measure results by these strata increases the ability of the reporting organizations to target areas for improvement without having to reverse engineer their measure results.

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

Based on the comparative analysis, the HCC-RRU approach (the variant of the CMS-HCC model) was noted as a viable alternative to the initial RRU risk adjustment approach. As with the ERG model, the HCC-RRU approach showed greater accuracy at the individual (member) level in predicting resource use, as indicated by individual r-squared analysis (see Table 2). The individual r-square statistic represents the percent of the variation across patients explained by a model; a higher r-square represents a more accurate model. The initial approach (Model 1) had an r-square of 5%, in contrast, the r-square for Model 4 is 48% (see Table 2). While Model 2 demonstrated some improvement in accuracy at the member level, the improvement was determined to be marginal; Model 2 was dropped from further analysis.

Table 2. Individual R-squared values; by Risk Adjustment Model Tested			
	Initial (Model 1)	Alternate (Model 2)	HCC-RRU (Model 4)
Medical Costs	0.050	0.081	0.482
Medical + Rx Costs	0.070	0.119	0.500

In the context of health plan measurement and their RRU result, NCQA additionally examined to what degree, if any, the improved precision of the HCC-RRU approach will impact health plans' O/E RRU results compared to the initial approach. As shown in Table 3, using the results for the 44 plans included in the research database, changing from the initial model to an approach based on the HCC-RRU model had a small to moderate impact on plans' O/E results. In general, the O/E results across plans were similar between Model 1 and Model 4, however some differences were observed for selected plans. While the difference in RRU ratio results is modest (approximately +/-5% on average) for the majority of the health plans tested, for some plans the difference in RRU result was more sizable (+/-15%).

Table 3: Absolute Difference in O/E Ratios Between Model 1 and Model 4- Medical Costs						
Condition	Mean	Min	P25	Median	P75	Max
Asthma	0.05	0.00	0.01	0.05	0.07	0.19
Cardiovascular	0.06	0.00	0.02	0.05	0.08	0.17
COPD	0.08	0.01	0.02	0.06	0.12	0.24
Diabetes	0.05	0.00	0.02	0.05	0.07	0.16
Hypertension	0.05	0.00	0.02	0.04	0.06	0.18

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*)

Approach to testing risk model was solely focused on appositeness testing for HEDIS RRU reporting as the chosen risk model was directly derived from the CMS-HCC approach. In the development of the CMS-HCC model, CMS evaluated several approaches that rely on diagnoses and ultimately selected CMS-HCC after determining it best met their criteria for health-based payment adjusters (transparency, ease of modification, and clinical coherence). In the CMS approach, each clinical category (CC) should contain relatively homogeneous diagnoses with respect to their expenditures. When hierarchies are applied, a patient is only coded for the most severe manifestation of their related disease and due to its reliance on specific coding, the hierarchy classifies vague diagnostic and lower-paying codes to lower categories thereby incentivizing the most specific coding possible.²

This approach has been extensively validated for its ability to balance expenditure predictions across differing populations and calibrated using a regression model of Medicare payment data.

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*):

N/A – the HCC-RRU risk model used for RRU reporting did not undergo additional statistical testing by NCQA. The NCQA model uses a selection of the risk weights provided by CMS.

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

N/A

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

N/A

2b4.9. Results of Risk Stratification Analysis:

N/A

² Pope GC et al. Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC model. Health Care Financing Review (25)4: 119-141, Summer 2004

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)

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*)

NCQA worked with the health plan field test sites who submitted blinded member-level data to NCQA along with the health plan's estimated risk weight following the HCC-RRU specification instructions. NCQA re-estimated these individual member risk weights based on the data submitted to NCQA. Finally, NCQA compared the plan estimated risk weights with those that were re-estimated. Additionally, NCQA administered a tracking survey at the beginning of the field test, requesting information about the feasibility and resource burden during their implementation of the HCC-RRU field test specification. Finally, NCQA posted the final field test HCC-RRU specification during a 30-day Public Comment period, available to all stakeholders, in July of 2009. These comments were reviewed and considered by NCQA and NCQA's Efficiency Measurement Advisory Panel (EMAP).

Of the three sites submitting data, one site matched NCQA's re-estimation of each member's risk weight exactly, the other sites had estimated risk weight mismatch occurring for 14% and 23% of the members respectively. Looking at the eligible populations separately, we found the number of member's not matching the NCQA risk weight estimate was approximately evenly distributed.

There was a substantial amount of variation in the time required for programming, ranging from 16-200 hours. Furthermore, health plans reported substantial variation in the amount of staff hours typically required to program any new HEDIS measure (not just the RRU measures), ranging from 48-160 hours. The reported time for the new HCC-RRU programming and any new HEDIS measure was not substantially different. Plans also reported between 2-10 hours to check for accuracy of the programming. For data collection, health plan sites reported staff resources between 2-87 hours to run the program and produce the field test data submission file, with 2-8 hours of that time to check for accuracy. Finally, during the actual submission process, field test sites reported an estimated 1 hour for submission and 1 hour for an accuracy review of the submission file.

As part of the survey, NCQA asked plans to report any significant obstacles to implementing the HCC-RRU risk adjustment model; all reported that that did not identify any obstacles to following and completing the HCC-RRU specification.

The feasibility and burden of implementing the refined HCC-RRU risk adjustment approach as part of the NCQA HEDIS RRU measures was examined extensively. Overall, it was found that this refined approach would be feasible for plans to implement. The burden associated with this more complex specification varied across plans, appearing to vary depending on their database environment. NCQA is determined to be sensitive and proactive in managing the burden associated with this more complex approach. NCQA has worked extensively with experts to include or expand on instructions in the specifications that address obstacles for implementation, clarify specifications to be applicable across environments, and have included both an HCC-RRU SAS pseudo-code and a full XML measure logic and reporting schema as part of the regular specification releases.

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)

The investigations described in attachment “SA_Reliability_VValidity Testing” provided insights into the conceptual and methodological issues in measuring relative utilization at a health plan level.

NCQA also performed detailed analysis on the most recent data available (2012) to discern the extent to which the relative resource use results reflect or express meaningful differences in performance. Bootstrap standard errors estimated for a given eligible population size multiplied by the z-value corresponding to a two-sided 95% confidence interval ($z=1.96$) were calculated and the results are shown below in Section 2b.5.2.

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)

NCQA used an INOVALON calculated typical standard error for specified eligible population sizes (This is the mean of bootstrap standard errors estimated for all 61 Markets); NCQA then derived a formula that uses these bootstrap standard errors and sample sizes to interpolate standard errors for health plan submissions with a given eligible population size. We then use these values to calculate the relative margin of error for each plan’s O/E ratio. The analysis was conducted separately for each product line, however HMOs and PPOs were combined together within product lines.

Tables 4 and 5 display the precision of the estimated O/E ratios

Table 4. Summary of the Relative Margin of Error for O/E Ratios from 61 Markets							
Resource Use Type	Cumulative Distribution of Plans by Margin Category						
	Total Count	Count with Valid O/E Ratios	Margin of Error (as a % of Estimated O/E Ratio)				
			≤ 5%	≤ 10%	≤ 15%	≤ 20%	> 20%
Total Medical	61	60	43.3	55.0	83.3	95.0	100.0
Total Pharmacy	61	60	43.3	75.0	95.0	100.0	100.0

Table 5. Summary of the Relative Margin of Error for O/E Ratios from Health Plan Submissions by Product Line								
Product Line	Resource Use Type	Cumulative Distribution of Plans by Margin Category						
		Total Count	Count w/Valid O/E Ratios	Margin of Error (as a % of Estimated O/E Ratio)				
				≤ 5%	≤ 10%	≤ 15%	≤ 20%	> 20%
Commercial	Total Medical	349	344	36.0	64.2	79.9	87.5	100.0

	Total Pharmacy	349	344	42.4	71.8	82.8	91.0	100.0
Medicaid	Total Medical	90	75	9.3	42.7	62.7	70.7	100.0
	Total Pharmacy	90	76	15.8	51.3	69.7	76.3	100.0
Medicare	Total Medical	204	187	28.9	61.0	78.6	87.2	100.0
	Total Pharmacy	204	187	34.8	64.7	82.4	88.2	100.0

The primary use of RRU O/E ratios is to determine if a health plan's predicted resource use was significantly different from the resource use we'd expect given the health plan's mix of patients. For Tables 6 and 7, NCQA classified the magnitude of the O/E ratio. Ratios between 0.95 and 1.05 are "<5%"; ratios of 1.05 to <1.1 or 0.95 to > 0.90 are "≥ 5%"; ratios of 1.10 to <1.15 or 0.90 to > 0.85 are ≥ "10%"; ratios of 1.15 to <1.20 or 0.85 to > 0.80 are "≥ 15%"; and ratios ≥ 1.20 or ≤ 0.80 are "≥ 20%". We calculated the percent of plans in each magnitude category. The denominator for the percentages is the "Count of Plans" in the same row. For example, 24.7% of plans with an O/E ratio that was significantly lower than 1.0 used between 5% and less than 10% fewer resources than expected. Table 6 shows results for Total Medical and Table 7 shows results for Total Pharmacy.

Table 6. Percent of Health Plans (within significance status1) by Magnitude of the Total Medical O/E Ratio								
Product Line	Significance of O/E Ratio	Count of Plans	Missing Ratio	Percentage Above or Below Expected				
				< 5%	≥ 5%	≥ 10%	≥ 15%	≥ 20%
Commercial	Missing	5	100.0	0.0	0.0	0.0	0.0	0.0
	Less than 1.0	77	0.0	10.4	24.7	32.5	14.3	18.2
	Not different	205	0.0	66.3	18.5	6.8	5.4	2.9
	Higher than 1.0	62	0.0	16.1	25.8	22.6	24.2	11.3
Medicaid	Missing	15	100.0	0.0	0.0	0.0	0.0	0.0
	Less than 1.0	12	0.0	0.0	8.3	41.7	33.3	16.7
	Not different	52	0.0	46.2	23.1	13.5	3.8	13.5
	Higher than 1.0	11	0.0	9.1	9.1	27.3	18.2	36.4
Medicare	Missing	17	100.0	0.0	0.0	0.0	0.0	0.0
	Less than 1.0	38	0.0	0.0	18.4	23.7	18.4	39.5
	Not different	91	0.0	54.9	19.8	16.5	5.5	3.3
	Higher than 1.0	58	0.0	3.4	13.8	22.4	22.4	37.9

Table 7. Percent of Health Plans (within significance status1) by Magnitude of the Total Pharmacy O/E Ratio								
Product Line	Significance of O/E Ratio	Count of Plans	Missing Ratio	Percentage Above or Below Expected				
				< 5%	≥ 5%	≥ 10%	≥ 15%	≥ 20%
Commercial	Missing	5	100.0	0.0	0.0	0.0	0.0	0.0
	Less than 1.0	105	0.0	8.6	21.0	17.1	17.1	36.2
	Not different	142	0.0	67.6	21.1	7.7	2.1	1.4
	Higher than 1.0	97	0.0	10.3	28.9	27.8	18.6	14.4
Medicaid	Missing	14	100.0	0.0	0.0	0.0	0.0	0.0
	Less than 1.0	32	0.0	0.0	9.4	15.6	12.5	62.5
	Not different	30	0.0	40.0	23.3	10.0	23.3	3.3
	Higher than 1.0	14	0.0	0.0	0.0	64.3	35.7	0.0
Medicare	Missing	17	100.0	0.0	0.0	0.0	0.0	0.0
	Less than 1.0	92	0.0	0.0	9.8	19.6	23.9	46.7
	Not different	48	0.0	54.2	25.0	6.3	10.4	4.2
	Higher than 1.0	47	0.0	10.6	8.5	63.8	6.4	10.6

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?*) Results of the sixth year analyses (2012 HEDIS data) of the RRU measurement set presented in Section 2b5.2 above illustrate the following:

- The standard errors for the Total Medical cost component are higher and this is evident in the lower proportion of plans that have a relative margin of error < 10%.

- The margin of error heavily influenced by sample size. 46 of the “Markets” in this analysis had eligible population sizes of at least 400 members. The higher margins of error were almost exclusively observed in “Markets” with fewer than 400 members.
- The standard errors for the Total Medical cost component are higher and this is evident in the lower proportion of plans that have a relative margin of error < 10%.
- The standard error of O/E ratios for Total Pharmacy is lower than for Total Medical. Therefore, we do see more plans demonstrating significant differences from 1.0 even when those ratios are closer to 1.0. This is result of the higher precision in Total Pharmacy O/E ratios.
- Regardless of product line and reporting type the majority of plans that were significantly different from a ratio of 1.0 used at least 10% fewer or greater resources than expected.
- Most plans that did not have an O/E ratio significantly different from 1.0 demonstrated resource use within 10% higher or lower than expected.
- This may indicate a convenient effect size. However, further study is warranted in order to determine if a difference of 10% is meaningful.

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*)

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*)

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*)

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)

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)

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*)