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:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☒ Cost/resource	☐ Process
☐ Efficiency	☐ 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 specificaitons** (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 (*incuding 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:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ 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:

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

1.3. What are the dates of the data used in testing? 2003-2012

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:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other: Click here to describe	☐ 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).

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

The 2012 HEDIS RRU data reports relative resource use for approximately 4,955,438 males (3,445,096 commercial; 199,172 Medicaid and 1,311,169 Medicare) and 4,567,115 females (2,870,298 commercial; 381,760 Medicaid and 1,315,057 Medicare) patients aged 42 years and above with COPD.

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. N/A

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 COPD* 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 COPD. 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.*

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.

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*)

We have included results from our initial Development Field Test, our Implementation and Feasibility Field test and our Annual Report of HEDIS RRU Submissions

Development Field Test Results: For the detailed analyses and results of the RRU Development Field test please refer to section 2b2.3 and attachment *SA_Reliability_Validity+Testing.pdf.*

Data Element Reliability: Overall, for imaging claims, the percentage of dollars with acceptable coding ranged from 80.3% for 2 records with a different CPT to 99.5% for 2 records with one CPT/one revenue code. Looking across plans for each combination showed reasonable consistency in percentage of dollars within each of the combinations. We reviewed the total records "% of Dollars with Acceptable Coding" amounts, by health plan and measured key percentiles to assess variation. Although there may be some variation in the relative records and

dollars across the combination scenarios, the variation in the "% of Dollars with Acceptable Coding" across plans was minimal.

There are often multiple tests included in a lab panel and we needed to know whether the codes were specific as to which tests are included in a panel with enough reliability to price the panel compared to the individual corresponding tests. An analysis of 18.9 million claim records for these services showed that results did not differ across health plans therefore inaccurate coding of service quantity for labs does not present an issue that would prevent reliable pricing.

In order to determine if a pricing strategy would be consistent for revenue codes vs. the CPT codes for similar services, we investigated how much variation would exist in observed payments within a revenue code. Variation was observed across records for both revenue and CPT coded services although greater variation was observed for revenue codes. The final results of this analysis[1], indicated 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.
- For many lab-related CPT codes, modifiers are used sparsely given that professional services are not expected to be utilized

Annual Analysis of RRU HEDIS Submissions: NCQA sets specific objectives for the RRU Annual analysis in order to examine the continued reliability and validity of the RRU HEDIS data supporting the measures:
Objective: Are a sufficient number of plans reporting RRU data? If too few, our estimates of expected resource use may be unreliable.

> Results: For the most recent year (2012), 669 plans reported audited, validated resource use data for COPD. Health plans were excluded due to extreme outlier errors (n=13) or too few members to measure reliably (n=192)

Objective: Are plans' observed-to-expected results for the RRU measures stable over time? If O/E changes too much year over year, this could indicate unreliability of the metric.

> Results: An indicator of plan stability over time is quartile movement of O/E ratios (for specific and overall service categories), with plans that move less than one quartile being considered stable, with the magnitude of absolute change being more relevant as opposed to the direction of change (up or down). 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 (Table 1 and Figures 1 & 2).

[1] Additional details from this study can be found in Attachment *SA_Standardized_Price_Implementation.pdf*

Table 1: Proportion of Plans with O/E Ratios that Changed by At Most One Quartile between Successive Years – 2012 v. 2011 (COPD)

	Percent of Plans with no more than 1 quartile shift (2012 vs. 2011)							
	HMOs				PPOs			
	Total Medical		Total Pharmacy		Total Medical		Total Pharmacy	
Product Line	Plan Count	% of Plans	Plan Count	% of Plans	Plan Count	% of Plans	Plan Count	% of Plans
Commercial	56	85.7	63	90.5	52	67.3	53	88.7
Medicaid	35	77.1	43	83.7	0	0	0	0
Medicare	60	83.3	67	92.5	8	100	8	100

Objective: What is the precision of the O/E ratios estimated for plans reporting data to NCQA: Imprecise estimates may suggest reliability problems?

Results: In terms of O/E ratio outlier distribution for the COPD RRU measure, 0.5% of commercial HMO and PPO plans were eliminated from *Total Medical* O/E results falling outside the pre-defined outlier range. In the Medicaid reporting line, approximately 1.1% of plans were found to have *Total Medical* O/E results below the 0.333 outlier threshold and 2.3% of plans were found to have the *Total Medical* O/E results above the 3.0 threshold. For Medicare, approximately 0.7% of plans were found to have *Total Medical* O/E results below the 0.333 outlier threshold and 2.0% of plans were found to have the *Total Medical* O/E results above the 3.0 threshold.

Figure 1: Distribution of *RCO Total Medical* O/E ratios (2012)

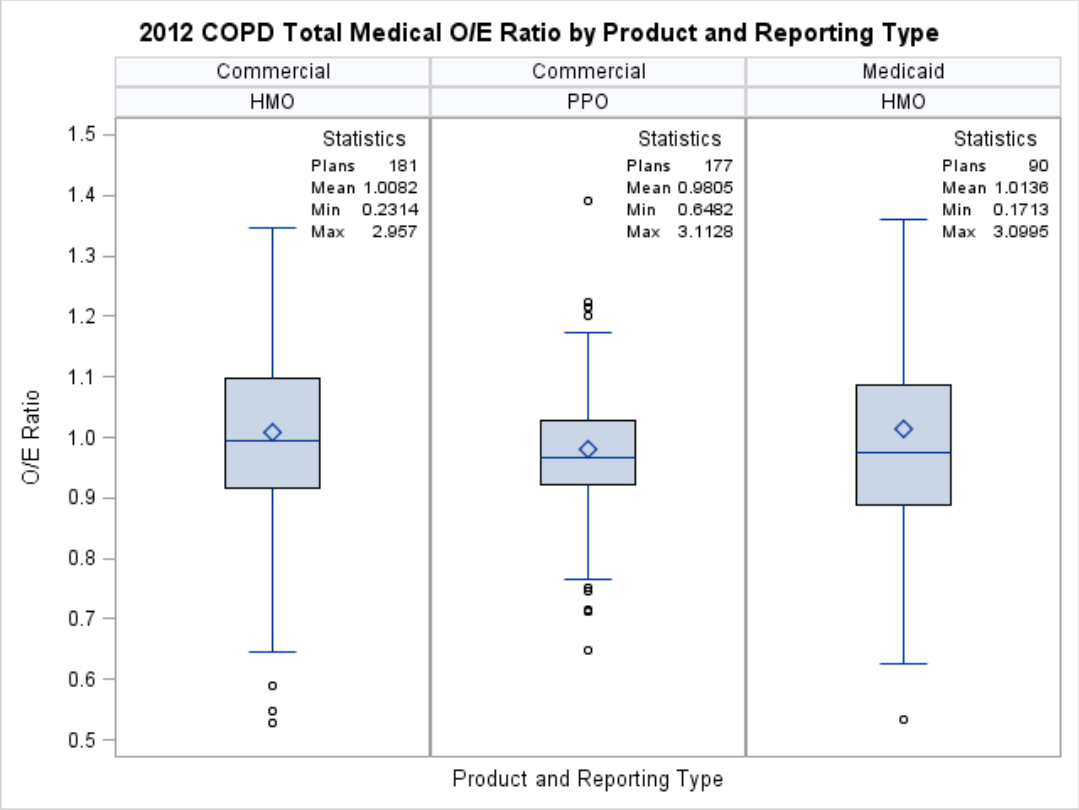
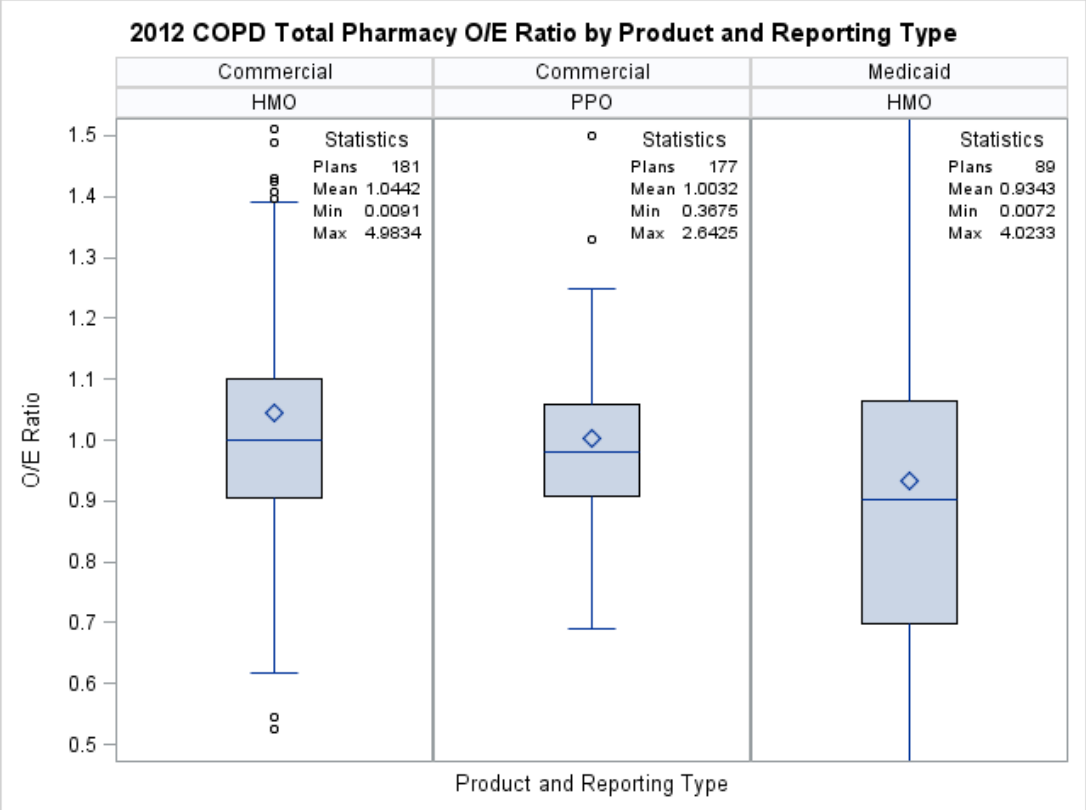


Figure 2: Distribution of RCO Total Pharmacy O/E ratios (2012)



Objective: Are correlations evident between the cost and quality components of the RRU measures? What is the strength and consistency of the association (if evident)?

Results: Component-component correlations generally provide a sense of the consistency of associations between RRU cost components (e.g., Total Discharges and Inpatient Facility) within each measure (by product line and reporting type) from year to year. 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). In 2012 new correlations considered 'moderate to strong positive' emerged across all product line-reporting type combinations. For additional details about the emerging component correlations refer to Tables A-8a to A12e in the Appendix of Attachment *2013 Analytic Report.pdf*.

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?*)

Development Field Test: For a more detailed interpretation and results of the Development Field test please refer to section 2b2.3 and Attachment- *SA_Reliability_Validity+Testing.pdf*

Data Element Feasibility: For a more detailed interpretation and results of the Implementation Field test please refer to Attachment- *SA_Standardized_Price_Implementation.pdf*

RRU Annual Analysis Reliability: Results from the most recent annual analyses indicate that plan performance generally remained stable across the 2011-2012 reporting period. With respect to RRU-specific HEDIS data, a total of 428 HMO and 244 PPO plans submitted data which is approximately 82% of all commercial HMOs and 94% of all commercial PPOs, 48% of all Medicare HMOs and 48% of all Medicare PPOs, and 49% of all Medicaid HMOs. Overall 48-94% of health plans reporting any HEDIS also reported HEDIS RCO (commercial 82-93%, Medicare & Medicaid ~48%). This plus the fact that 81-96% of plans reporting RRU voluntarily chose to publically report the RCO measure indicates a high level of confidence by the plans that their data is a good representation of their relative resource use for the year. Correlation analyses demonstrated an increased precision of the updated risk adjustment model, with new positive and negative correlations emerging in 2012 and in term of plan stability, returning plans were more successful in terms of not being eliminated due to outlier distribution and data completeness for both the *Total Pharmacy* and *Total Medical* O/E ratios. A more complete discussion of the investigation into outlier distribution, data completeness and availability of data for public use can be found in the Attachment *2013 RRU Analytic Report.pdf* starting on p.12.

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.

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.

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 Validity[2]: 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 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.

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

The Development Field Test 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? (See Table 5, page 35 of Attachment *SA_Reliability_Validity+Testing.pdf*):

- 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? (See Table 6, page 36 of Attachment *SA_Reliability_Validity+Testing.pdf*):

- 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? (See Tables 7 & 8, pp. 38-40 of Attachment *SA_Reliability_Validity+Testing.pdf*):

[2] 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*

- 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.

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 Attachment *SA_Reliability_Validity+Testing.pdf*). 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. 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.

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. 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. The Development Field Test 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.
- The Development Field Test 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.

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 both product lines who had evidence of other dominant medical conditions, such as active cancer, specific organ transplants (non-renal), HIV/AIDS, ESRD and organ transplants are required to be excluded from RRU measurement due to the excessive costs associated with treatment for these conditions.

Cost-related Exclusion Testing: 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 segmentation of the RRU research database by disease and risk cohorts to look for 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) both were major contributors to costs incurred regardless of the primary condition or service category in consideration, as seen in Table 2. Additionally, the largest proportion of Total Medical costs was associated with the Inpatient Facility service sub-category.

Table 2: Relative Cost (PMPM) Ratio of Cohort 4 to Cohort 3, COPD

Age	Gender	# Members	Member Months		Evaluation & Management		Procedure & Surgery		Total Pharmacy	INPT Facility	Imaging	Lab	Total; Medical	Total
			Medical	Pharmacy	INPT	OUTPT	INPT	OUTPT						
18-44	**F**	19	227	167	6.5	1.4	6.4	3.9	3.8	7.1	1.7	3.8	5.1	4.8
18-44	**M**	17	204	84	10.6	2.1	8.2	6.2	4.5	12.0	2.6	4.9	8.6	7.7
45-54	**F**	176	2,112	1,248	7.1	1.5	4.9	2.9	2.3	7.4	1.9	3.3	5.2	4.5
45-54	**M**	212	2,541	1,641	6.8	1.7	5.3	4.0	2.6	7.1	2.5	4.0	5.6	4.9
55-64	**F**	456	5,470	3,082	6.3	1.4	4.8	3.1	1.8	6.5	1.7	3.1	4.8	4.0
55-64	**M**	617	7,399	4,473	5.7	1.7	4.2	3.0	2.3	5.5	2.0	3.0	4.4	3.9
	total	1,497	17,953	10,695	6.38	1.58	4.83	3.23	2.25	6.46	1.88	3.24	4.90	4.25

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 ESRD comorbidities in the COPD population is comparatively 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 standardized price 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 plan performance.

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 HHC-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 for comparison.

Stratification of RRU Results

NCQA summarizes resource measures for all 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. The HCC-RRU approach showed greatest accuracy at the individual (member) level in predicting resource use, as indicated by individual r-squared analysis (Table 4). 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%.

Table 4. Individual R-squared values; by Risk Adjustment Model Tested

	Model 1	Model 2	HCC-RRU
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 5, 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 HCC-RRU, 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 5: Absolute Difference in O/E Ratios Between Model 1 and HCC-RRU- 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.[3] This approach has been

[3] 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

extensively validated for its ability to balance expenditure predictions across differing populations and calibrated using a regression model of Medicare payment data.[4]

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*)**:**

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

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*)

In order to test the feasibility of reporting the HCC-RRU risk model, NCQA worked with 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.

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.

Overall, the feasibility and burden of implementing the refined HCC-RRU risk adjustment approach was found to feasible for plans to implement. The burden associated with this more complex specification varied across plans, appearing to vary depending on their database environment.

[4] Evaluation of the CMS-HCC Risk Adjustment Model: Final Report. Centers for Medicare & Medicaid Services. March 2011

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)*

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. Boot strap 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*)

In order to investigate the precision of estimated O/E ratios, OptumInsight provided NCQA with the typical standard error for specified eligible population sizes using the mean of bootstrap standard errors estimated for 61 markets. [15] NCQA then derived a formula using these standard errors and sample sizes to interpolate standard errors for health plan submissions with a given eligible population size. These were then multiplied by the z-value corresponding to a two-sided 95% confidence interval (z = 1.96) to yields the absolute margin of error, which are then presented as a proportion of the estimated O/E ratio (Tables 6 & 7). These values are then used to calculate the relative margin of error for each plan's O/E ratio. The relative margin of error indicates the reliability of the estimate to within a specified percentage above or below the point estimate of the O/E ratio. The relative margins of error are divided into five categories: ≤ 5%, ≤ 10%, ≤ 15%, ≤ 20%, and > 20% and the percent of plans falling into each category was reported (separately) for the Total Medical and Total Pharmacy cost components. Table 6 exhibits the results of this analysis using bootstrap standard errors.

Table 6. Summary of the Relative Margin of Error for O/E Ratios from 61 Markets - (COPD)

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	60	60	36.7	45.0	80.0	95.0	100.0
Total Pharmacy	60	60	38.3	61.7	88.3	96.7	100.0

The standard errors for the Total Medical cost component are higher which is evident in the lower proportion of plans that have a relative margin of error < 10%. The margin of error is heavily influenced by sample size. Smaller markets have generally have larger margins of error. Table 7 shows the same strategy applied to the actual health plan submissions for RCO stratified by product line.

[5] Markets" define combinations of data contributor (e.g., health plan) and geography

Table 7. Summary of the Relative Margin of Error for O/E Ratios from Health Plan Submissions by Product Line - (COPD)

Product Line	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%
Commercial	Total Medical	350	347	35.7	63.4	79.3	87.3	100.0
	Total Pharmacy	350	346	41.9	70.8	85.5	92.5	100.0
Medicaid	Total Medical	94	88	34.1	58.0	76.1	84.1	100.0
	Total Pharmacy	94	88	38.6	65.9	77.3	85.2	100.0
Medicare	Total Medical	209	197	49.7	75.6	89.3	95.4	100.0
	Total Pharmacy	209	197	54.8	81.2	91.4	96.4	100.0

Compared to Inovalon's data, health plans submissions from HEDIS tended to have smaller margins of error given the higher proportions of plans with margins of error of 10% or lower.

The primary use of O/E ratios for RRU 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. In Tables 8 & 9 below, NCQA classified the magnitude of the O/E ratio as follows: 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%". The percent of plans in each category is then determined and the denominator for the percentages becomes the "Count of Plans."

Table 8. Percent of Health Plans (within significance status1) by magnitude of the Total Medical O/E Ratio - (COPD)

Product Line	Significance of O/E Ratio	Count of Plans	Percentage Above or Below Expected				
			< 5%	≥ 5% to <10%	≥ 10% to <15%	≥ 15% to <20%	≥ 20%
Commercial	Missing	3	--	--	--	--	--
	Less than 1.0	89	12.4	24.7	27.0	11.2	24.7
	Not different	180	67.8	16.7	10.6	2.8	2.2
	Higher than 1.0	78	12.8	25.6	30.8	12.8	17.9
Medicaid	Missing	6	--	--	--	--	--
	Less than 1.0	27	0.0	22.2	29.6	22.2	25.9
	Not different	37	59.5	13.5	8.1	10.8	8.1

	Higher than 1.0	24	12.5	29.2	25.0	4.2	29.2
Medicare	Missing	12	--	--	--	--	--
	Less than 1.0	54	11.1	16.7	11.1	24.1	37.0
	Not different	51	66.7	19.6	11.8	0.0	2.0
	Higher than 1.0	92	4.3	19.6	12.0	13.0	51.1

Table 9. Percent of Health Plans (within significance status1) by magnitude of the Total Pharmacy O/E Ratio - (COPD)

Product Line	Significance of O/E Ratio	Count of Plans	Percentage Above or Below Expected				
			< 5%	≥ 5% to <10%	≥ 10% to <15%	≥ 15% to <20%	≥ 20%
Commercial	Missing	4	--	--	--	--	--
	Less than 1.0	114	7.9	28.9	21.9	14.0	27.2
	Not different	134	56.7	24.6	12.7	3.0	3.0
	Higher than 1.0	98	13.3	18.4	22.4	12.2	33.7
Medicaid	Missing	6	--	--	--	--	--
	Less than 1.0	44	2.3	11.4	13.6	9.1	63.6
	Not different	21	52.4	28.6	4.8	14.3	0.0
	Higher than 1.0	23	4.3	21.7	17.4	8.7	47.8
Medicare	Missing	12	--	--	--	--	--
	Less than 1.0	97	7.2	12.4	22.7	23.7	34.0
	Not different	34	67.6	20.6	5.9	2.9	2.9
	Higher than 1.0	66	3.0	18.2	13.6	12.1	53.0

These investigations based on the 2012 HEDIS submission led to the conclusions that, regardless of product line and reporting type, most plans that were significantly different from 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.

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.

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*)