NATIONAL QUALITY FORUM

TO: NQF Members and Public

FR: NQF Staff

RE: Pre-voting review for *National Voluntary Consensus Standards for Cost and Resource Use (Cycle 1): A Consensus Report*

DA: August 30, 2011

Resource use measures count the frequency of defined health system resources, are broadly applicable and comparable measures of health services counts that are applied to a population or event. This project seeks to endorse cost and resource use measures, which will serve as building blocks for efficiency of care measures and signal the measure development industry of the urgent need to develop measures of efficiency that integrate quality domains with cost and resource use measures. This is NQF's first effort focused on endorsing cost and resource use measures.

Four condition-focused Technical Advisory Panels (TAPs) for pulmonary, cardiovascular and diabetes, bone and joint, and cancer conditions were convened to assist the project's 23-member Steering Committee in making recommendations. In this first review cycle, seventeen measures were evaluated for suitability as voluntary consensus standards for accountability and performance improvement; of those, nine measures were withdrawn by the developer. To date, the Steering Committee has recommended four cost and resource use measures for endorsement.

The draft document, *National Voluntary Consensus Standards for Cost and Resource Use (Cycle 1): A Consensus Report* is posted on the NQF website along with the following additional information:

- measure submission forms; and
- meeting and call summaries from the TAP and Steering Committee's discussions.

In addition to commenting on the recommended measures, we request that comments also be submitted on the "Applying Resource Use Measure Evaluation Criteria" section of the report. Please submit your comments on this section in the "General Comment" section of the commenting tool, including a reference to the page and line number in the report.

Pursuant to section II.A of the Consensus Development Process v. 1.8, this draft document, along with the accompanying material, is being provided to you at this time for purposes of review and comment only and is not intended to be used for voting purposes. You may post your comments and view the comments of others on the NQF website.

NQF Member comments must be submitted no later than 6:00 pm ET, September 28, 2011.
Public comments must be submitted no later than 6:00 pm ET, September 21, 2011.

Thank you for your interest in NQF's work. We look forward to your review and comments.

NATIONAL QUALITY FORUM

NATIONAL VOLUNTARY CONSENSUS STANDARDS FOR COST AND RESOURCE USE (CYCLE 1): A CONSENSUS REPORT

DRAFT REPORT FOR COMMENTING

AUGUST 30, 2011

NATIONAL VOLUNTARY CONSENSUS STANDARDS FOR COST AND RESOURCE USE (CYCLE 1): A CONSENSUS REPORT

TABLE OF CONTENTS

NATIONAL VOLUNTARY CONSENSUS STANDARDS FOR COST AND RESOURCE USE (CYCLE 1): A CONSENSUS REPORT

EXECUTIVE SUMMARY

As current health reform efforts focus on expanding coverage, increasing access to care, and reducing costs, it is important to understand how the system uses resources in the context of health outcomes. Combining resource use (or cost) and quality data will enable the system to better evaluate efficiency of care. Understanding resource use measurement as a building block of efficiency is a first step toward this goal. For the purposes of this project, resource use measures are defined as broadly applicable and comparable measures of health services counts (in terms of units or dollars) that are applied to a population or event (e.g., diagnoses, procedures, or encounters). A resource use measure counts the frequency of defined health system resources; some may further apply a dollar amount (e.g., allowable charges, paid amounts, or standardized prices) to each unit of resource use.

This Consensus Development Process (CDP) project will endorse resource use (or cost) measures that will serve as building blocks for efficiency of care measures and signal the measure development industry of the urgent need to develop resource use and efficiency that integrate quality domains with resource use measures. In applying the Resource Use Measure Evaluation Criteria for the first time, the Technical Advisory Panels (TAPs) and Steering Committee encountered several overarching issues during their discussions and evaluations of the measures. Some issues varied by developer as each developer submitted measures with very distinct approaches. This report reflects the discussion of those issues as well as the measure-specific evaluation summaries for four measures reviewed during the first review cycle. A subsequent report will address the remaining measures considered during this project and any additional recommendations provided by the Committee.

Four measures are recommended for endorsement as voluntary consensus standards suitable for accountability and performance improvement:

- (1557) Relative Resource Use for People with Diabetes (RDI) (NCQA)
- (1558) Relative Resource Use for People with Cardiovascular Conditions (NCQA)
- (1598) Total Resource Use Population-based PMPM Index (HealthPartners)
- (1604) Total Cost of Care Population-based PMPM Index (HealthPartners)

NATIONAL VOLUNTARY CONSENSUS STANDARDS FOR COST AND RESOURCE USE (CYCLE 1): A CONSENSUS REPORT

BACKGROUND

The United States' health care expenditures are unmatched by any country in the world. [1] This spending, however, has not resulted in better health for Americans. In fact, higher spending has not led to lower mortality, greater patient satisfaction, improvements in access to health care, or higher quality care.[2,3,4] This phenomenon of high spending with disproportionate outcomes points to a system laden with waste. The contributing factors to this alarming trend are as complex as the health care system itself, with physician practice patterns, regional market influences, and access to care as major players. Meanwhile, the United States' health care spending continues to increase at a rate of seven percent per year, and is largely focused on treating acute and chronic illness rather than on preventative care.[5]

As ongoing health reform efforts focus on expanding coverage, increasing access to care, and reducing costs, it is important to understand how resources are currently being used in the system in the context of quality, preferably related to health outcomes. The combination of resource use (or cost) and patient quality data will enable the system to better evaluate efficiency of care. Several provisions in the Affordable Care Act (ACA), slated to be implemented over the next five years, require using resource use data to further support efforts to move toward a value-based purchasing (VBP) payment model. One such provision requires the Secretary of Health and Human Services to develop an episode grouper that combines separate but clinically related items and services into an episode of care for an individual.[6] Additionally, resource use data will also be included on the physician compare website, as well as a physician value modifier that will be used to adjust fee-for-service (FFS) payments by combining physician performance on quality and resources use. While the ACA legislation is focused on the Medicare population, understanding resource use measurement as a building block of efficiency, even in the context of commercial-based measures, is a first step toward meeting these goals.

For the purposes of this project, resource use measures are defined as broadly applicable and comparable measures of health services counts (in terms of units or dollars) that are applied to a population or event (broadly defined to include diagnoses, procedures, or encounters). A resource use measure counts the frequency of defined health system resources; some may further apply a dollar amount (e.g., allowable charges, paid amounts, or standardized prices) to each unit of resource use. Current approaches for measuring resource use range from broadly focused measures, such as per capita measures, which address total healthcare spending (or resource use) per person, to those with a more narrow focus, such as measures dealing with the healthcare spending or resource use of an individual procedure (e.g., a hip replacement).

This Consensus Development Process (CDP) project, the second phase of a two-phase effort, will endorse resource use measures, which will serve as building blocks for efficiency of care measures and signal the measure development industry of the urgent need to develop resource use and efficiency measures that integrate quality domains. Phase one, which began in 2009, was aimed at understanding resource use measures and identifying the important attributes to consider in their evaluation. During this phase, the current NQF Measure Evaluation Criteria used for the evaluation of quality measures was reviewed and refined by the Resource Use Steering Committee to address the unique aspects of resource use measures, resulting in the NQF Resource Use Measure Evaluation Criteria. A single Steering Committee was used across both phases of work, with the addition of four Technical Advisory Panels (TAPs) in phase two to assist the Committee in evaluating the measures' clinical and methodological aspects. The CDP project was divided into two review Cycles between which fourteen focus areas were assigned:

Cycle 1	Cycle 2
Cardiovascular • Congestive heart failure (CHF) • Coronary artery disease (CAD) • Acute myocardial infarction (AMI)	**Pulmonary** • Chronic obstructive pulmonary disease (COPD) • Asthma • Pneumonia
Stroke	**Cancer** • Breast cancer • Colorectal cancer
Diabetes	
Non-condition specific (e.g. per capita-population)	**Bone/Joint** • Hip or knee replacement • Hip or pelvic fracture • Low back pain

This report reflects the discussion and overarching issues the Committee identified while evaluating cost and resource use measures submitted to the project; measure-specific evaluation summaries are provided only for a subset of Cycle one measures. A subsequent report will address remaining Cycle one measures as well as all Cycle two measures.

STRATEGIC DIRECTIONS FOR NQF

NQF's mission includes three parts: 1) building consensus on national priorities and goals for performance improvement and working in partnership to achieve them; 2) endorsing national consensus standards for measuring and publicly reporting on performance; and 3) promoting the attainment of national goals through education and outreach programs. As greater numbers of quality measures are developed and brought to NQF for consideration of endorsement, NQF must assist stakeholders in measuring "what makes a difference" and addressing what is important to achieve the best outcomes for patients and populations.

Several strategic issues have been identified to guide consideration of candidate consensus standards:

DRIVE TOWARD HIGH PERFORMANCE. Over time, the bar of performance expectations should be raised to encourage achievement of higher levels of system performance.

EMPHASIZE COMPOSITES. Composite measures provide much-needed summary information pertaining to multiple dimensions of performance and are more comprehensible to patients and consumers.

MOVE TOWARD OUTCOME MEASUREMENT**.** Outcome measures provide information of keen interest to consumers and purchasers, and when coupled with healthcare process measures, they provide useful and actionable information to providers. Outcome measures also focus attention on much-needed system-level improvements because achieving the best patient outcomes often requires a carefully designed care process, teamwork, and coordinated action on the part of many providers.

CONSIDER DISPARITIES IN ALL WE DO. Some of the greatest performance gaps relate to care of minority populations. Particular attention should be focused on identifying disparities-sensitive performance measures and on identifying the most relevant race/ethnicity/language/socioeconomic strata for reporting purposes.

NATIONAL PRIORITIES PARTNERSHIP AND THE NATIONAL QUALITY STRATEGY

The National Priorities Partnership, a multi-stakeholder collaborative of 48 organizations convened by NQF, plays a key role in identifying strategies for achieving national goals for quality healthcare and facilitating coordinated, multi-stakeholder action. The Department of Health and Human Services has asked the Partnership for its collective, multi-stakeholder input on the National Quality Strategy (NQS) framework, which includes three inextricably linked domains—better care, affordable care, and healthy people/healthy communities—around which priorities, goals, measures, and strategic opportunities for improvement are to be identified and/or refined.

When the NQS was announced in March 2011, one of the priorities it identified was Making Quality Care More Affordable. The resource use measure endorsement process is an important step toward measuring affordable care by evaluating resource use and cost measures. These measures can identify opportunities to reduce the rate of growth in health care spending, and when paired with quality measures, can help evaluate the efficiency of the health care system.

RELATED NQF WORK

This project is NQF's first effort focused on evaluating and endorsing cost and resource use measures. In 2009, NQF completed a measurement framework for evaluating efficiency across patient-focused episodes of care. This report, NQF Measurement Framework: Evaluating Efficiency across Patient-Focused Episodes of Care, presents the NQF-endorsed® measurement framework for assessing efficiency, and ultimately value, associated with the care over the course of an episode of illness and sets forth a vision to guide ongoing and future efforts.

RESOURCE USE MEASURES IN CONTEXT

This consensus development process seeks to endorse resource use (or cost) measures as building blocks toward measuring efficiency of care. Efficiency can be defined broadly as the resource use (or cost) associated with a specific level of performance with respect to the other five Institute of Medicine (IOM) aims of quality: safety, timeliness, effectiveness, equity, and patient-centeredness.[7] Resource use measures can also be used to assess value by integrating preference-weighted assessments of the quality and cost performance of a specified stakeholder, such as an individual patient, consumer organization, payer, provider, government, or society.[8]

As a building block in understanding efficiency and value, NQF supports the using and reporting of resource use measures in the context of quality performance, preferably outcome measures. Using resource use measures independent of quality measures does not provide an accurate

assessment of efficiency or value, and may lead to adverse unintended consequences in the health care system.

Resource use measures used to assess efficiency and value should be important to measure, have scientifically acceptable properties, and be usable and feasible. Those resource use measures under evaluation in this process should independently meet these endorsement standards. Future efforts will need to evaluate how resource use measures can be paired with appropriate quality measures to assess the healthcare system's efficiency. These efforts should consider quality and resource measure alignment of the underlying population, exclusions, and risk-adjustment, among other measure properties.

NQF'S CONSENSUS DEVELOPMENT PROCESS

NQF's National Voluntary Consensus Standards for Cost and Resource Use project seeks to endorse resource use and cost measures for performance improvement and accountability in the context of quality measures.

Evaluating Potential Consensus Standards

Candidate consensus standards were solicited through a Call for Measures on January 31, 2011. In the first review cycle, seventeen measures were submitted and evaluated for suitability as voluntary consensus standards for accountability; nine of these were withdrawn by the developer. The measures were evaluated using NQF Resource Use Measure Evaluation Criteria. Four condition-focused TAPs for pulmonary, cardiovascular and diabetes, bone and joint, and cancer conditions rated each candidate consensus standard according to the subcriteria and identified strengths and weaknesses to assist the Committee in making recommendations. The 23-member, multi-stakeholder Committee evaluated the subcriteria of the non-condition specific measures, provided final evaluations of the four main criteria—importance to measure and report, scientific acceptability of the measure properties, usability, and feasibility—and made

endorsement recommendations for all measures. Measure developers were available during TAP and Committee discussions to respond to questions and clarify any issues or concerns.

Principles for Resource Use Measure Evaluation

In Phase one of this project, the Committee defined resource use measures and their constructs to better understand how to evaluate these measures. For the purposes of this project, resource use measures are defined as broadly applicable and comparable measures of health services counts (units or dollars) applied to a population or event (diagnoses, procedures, or encounters). Resource use measure scores may be expressed as counts, dollars, or even observed-to-expected ratios. The Committee developed the following principles to frame its subsequent effort to refine the evaluation criteria for resource use measures:

1. Efficiency is one of the Institute of Medicine (IOM) five quality aims and is a function of resource use and health outcomes: *Efficiency = fx(resource use, health outcomes)*
2. Resource use measures are the amount of resources used per population, episode, or procedure.
3. Resource use measures are an important building block for measures of efficiency of care; future measurement efforts should integrate and explicitly incorporate measures of quality, health outcomes, or appropriateness.
4. The justification for and intended purpose of resource use measures is to examine, understand, and ultimately reduce unnecessary costs in care.
5. There is a continuum of resource use measures (i.e., per capita to per procedure); all types under consideration for endorsement must meet NQF evaluation criteria for such measures.
6. The resource use measure specification and calculation must be explicitly stated and transparent so the approach can be deconstructed and implemented in a standard manner.
7. Comprehensive measures are preferable, even if combining multiple service categories into one resource use estimate increases complexity; using methodologically sound methods is of paramount importance.
8. The final resource use measure or result should be simple and easy for all stakeholders to interpret.

9. Methods for combining the component scores influence the interpretation of the measure results and must be justified (e.g., averaging across all component scores may obscure low or high scores of individual components).
10. While resource use measure developers may have fundamental differences in approach, these principles should apply across all types and approaches.
11. NQF considers transparency as key to ensuring the intended audiences understand the results and can use them for decision making. Resource use measures are often highly complex, with lengthy algorithm decision trees that can make clarity difficult, particularly when some components may be only be partially transparent to the user.

Applying the Resource Use Measure Evaluation Criteria

With a working definition of resource use measures and guiding principles in place, the Committee completed a detailed review of the standard NQF Measure Evaluation Criteria. This review resulted in the NQF Resource Use Measure Evaluation Criteria, based on the same four major criteria used to evaluate quality measures - importance, scientific acceptability, usability, and feasibility - with targeted changes to the subcriteria to address the unique attributes of resource use measures.

In applying the Resource Use Measure Evaluation Criteria for the first time, the TAPs and Committee encountered several overarching issues during their discussions and evaluations of the measures. Some issues varied by developer as each developer submitted measures with very distinct approaches. The Committee factored these issues into their ratings and recommendations for multiple measures, recognizing the need to balance the quantity and specificity of information required to adequately evaluate the measure and the burden on the developer to provide this information. These issues are included below in the discussion of each criterion, in addition to the summary provided of each individual measure in the evaluation summary table.

Importance

The importance criterion for resource use measures, like that for quality measures, is aimed at determining the extent to which the focus of the measure (e.g., hip fractures, coronary artery disease) is important to measure and report. For resource use measures, the developers were asked to demonstrate high impact by showing there is variation and opportunities for improvement in the delivery of care for the identified condition. The TAP concluded that the measures submitted were broad and inclusive of high impact conditions. Additional subcriteria were tailored specifically for resource use measures. These subcriteria included an evaluation of whether the intent of the measure had been clearly described and whether the resource use service categories selected to measure costs accurately reflected the intent and focus of the measure. All measure submissions were found to be important.

Scientific Acceptability

Similar to quality measures, evaluation of the scientific acceptability of resource use measures includes the reviewing of the measure's specifications, reliability and validity testing, and approach to addressing disparities. Within the reliability criterion, the completeness, repeatability of the specifications, and the adequacy of the reliability testing methodology and results are evaluated. Applying the validity criteria, the Committee was asked to determine whether the specifications reflected the intent of the measure and addressed those areas where there was variation, as demonstrated in importance. The validity criterion also includes an assessment of the adequacy of validity testing, exclusions, risk-adjustment, and the identification of meaningful differences.

Resource Use Specification Modules

The resource use measure specifications were delineated by five main modules, including: 1) data protocol, 2) measure clinical logic, 3) measure construction logic, 4) adjustments for comparability, and 5) measure reporting. To allow for user flexibility, the developers were permitted to submit measurement steps in the data protocol and reporting modules as specifications or guidelines, or to not submit instructions at all. Specifications are inherent

measure characteristics that must be fully implemented in order to obtain valid measure results. Guidelines, on the other hand, are suggested approaches from the developer on possible ways to implement these steps. Evaluation of resource use measure specifications proved to be the most intensive effort in the review process. The issues identified within each of the specification modules have been outlined below.

Data protocol

The data protocol module allows developers to submit instructions and analytic steps for cleaning or aggregating relevant data necessary to implement the specifications and produce valid results. Measure developers submitted the following data protocol information: data preparation, data inclusion criteria, data exclusion criteria and considerations for missing data. Recognizing that not all developers create specifications around these steps, the Committee concluded these items could be submitted as specifications or guidelines, or not submitted at all.

All of the measures submitted use administrative claims as the data source. Administrative claims offer the benefit of reduced administrative burden for providers and measure implementers in collecting and reporting data elements. However, variation in coding practices has the potential to affect the reliability and validity of any measure that relies on administrative and claims data alone, including resource use measures. This may be particularly true for entities providing care under capitated financial arrangements that may capture fewer diagnostic and procedural codes per record than those operating under traditional FFS arrangements.

Accountable entities may outsource services through pharmacy benefit managers (PBMs) or behavioral/mental health carve-outs, which may result in incomplete or missing pharmacy or behavioral/mental health data. These entities can outsource administration of outpatient prescription drug benefits to PBMs.[9] Carve out arrangements allow accountable entities to separate behavioral/mental health insurance benefits by contracting with a third party to manage care and/or the insurance risk for patients requiring these services.[10] The Committee agreed that total resource use for entities that do not receive member claim information from carve-out

pharmacy and behavioral/mental health services may not be comparable to resource use for those that do not outsource these services. In this instance, interpreting the overall costs for a patient across health plans with and without carve-out arrangements would be misleading.

However, entities without member claims data from their carve-out arrangements can be flagged for comparison with entities with similar missing benefit information. Because resource use measures allow claims to be assigned to resource use categories (i.e. laboratory and imaging), these categories can be used to compare costs across entities even when outsourcing arrangements are present. For example, comparing laboratory costs or imaging costs across entities within a total per-capita resource use measure would be informative even when pharmacy data are not available.

Clinical logic

Evaluation of the measure clinical logic included steps to identify the condition or event of interest and any clustering of diagnoses or procedures. This evaluation included examining the clinical topic area and determining whether or not the measure accounts for co-morbid conditions, disease interactions, clinical hierarchies, clinical severity levels, and concurrency of clinical events.

The complexity of the submitted measure specifications made evaluating the measure's clinical logic challenging. For example, measure developers designed various methodologies to assign patients to a severity level; however, due to complex algorithms, specific details and code lists used to determine the assignment of patients to severity categories were difficult to interpret.

Exclusions were a focus during evaluation of the resource use measure's clinical logic. Although the creation of homogenous populations enables comparability, measure developers should ensure that measure exclusions do not allow for complications from poor care to drive patients out of the episode, thus rewarding entities that provide inadequate care. For example, a biased

measure score may be created by excluding patients with acute myocardial infarction (AMI) who are discharged from a skilled nursing facility or excluding patients who are not discharged alive.

Finally, resource use measures that seek to create more homogenous patient populations are often limited by the ability of administrative claims data to accurately assess patient health status and severity. For example, measure submitted did not have the ability to differentiate between community-acquired and healthcare-acquired pneumonia. Measures submitted also were not able to identify staging information to assess the severity of a cancer diagnosis.

Construction logic

The measure construction logic evaluation included a review of the steps used to cluster, group, or assign claims beyond those associated with the measure's clinical logic and an assessment of how the various components of the measure (episode logic, clinical logic, risk-adjustment) work together. Measures were evaluated to determine if the temporal parameters including trigger and termination rules are appropriate for the clinical logic specified within the measure. For example, the Committee evaluated the post-hospitalization period in an episode of AMI to ensure it was appropriate for the measure's intent, level of analysis, attribution approach and statistical properties.

For measures that were specified at the individual or group practice level, the Committee was particularly interested in the reliability and validity testing. The Committee expected developers to demonstrate statistical differences at sample sizes that would be observed in individual and group practices. Further, attribution of the measure to the individual or group practice level was discussed at length, focusing on the appropriateness and generalizability of the attribution approach. While sample size and attribution could be submitted as guidelines, the Committee agreed that these testing results contribute to the measure's scientific acceptability at these levels of analysis.

Measures that were submitted as a part of an episode grouper were more difficult to evaluate since the assignment of claims into the episode, comorbidities and interactions, clinical hierarchies, and the handling of concurrent of clinical events were a function of a grouper system. Measures submitted to this project were evaluated as standalone measures of resource use; however, the construction logic within episode grouper-based approaches include claim assignment decisions, or tie-breaker logic, which is not always clear when evaluating single measures or resource use. Tie-breaker logic is a mechanism to determine how a claim or record is assigned to an episode if it is eligible for assignment to multiple episodes. For example, if a patient fills a prescription that could be mapped to multiple open episodes, tie breaking logic could be used to determine how this cost would be assigned. Additional work is needed to determine specific evaluation criteria for episode grouper systems.

Adjustments for comparability

A measure's result can be influenced by confounding external factors that can impact the measure score. Measure developers submitted steps to adjust the measure to increase comparability. These adjustments include risk-adjustment, stratification approach, and the costing method used within the measure.

Risk-adjustment methodologies varied considerably across measure developers. A combination of the complexity and a varying degree of transparency of the risk-adjustment approach made evaluating the methods challenging. The experts agreed that the details on the performance of risk models were vital to determining the model's adequacy; specifically, how the presence of certain claims drive categorization into different risk categories and the risk model's goodness-of-fit. Of the various methodologies reviewed, none were considered to be superior. A Society of Actuaries report shared with the Committee comparing various risk-adjustment methodologies [e.g., Hierarchical Clinical Categories (HCC), Adjusted Clinical Groups (ACG), Episode-risk-group (ERG)] was informative; however, more research and guidance on the appropriateness of the models for specific applications are needed.

Stratification can be a mechanism to create homogenous risk populations; however, similar to the concern that exclusions may remove patients out of an episode inappropriately, measure developers need to ensure that the risk stratification approach does not allow for complications from poor care to drive patients into a higher risk stratum, thus rewarding entities who provide inadequate care. For example, for patients with coronary artery disease (CAD), creating risk strata based on subsequent revascularization has this potential for adverse consequences.

The developers were asked to specify a costing method to apply to the measure. For the measures submitted, the costing approaches were either specified for the actual amount paid (i.e., cost of care measures) or for standardized prices (i.e., resource use measure). Standardized pricing allows users to compare the use and intensity of health services while holding actual paid amounts constant. The Committee was divided on the utility of cost of care measures, as both approaches could be appropriate for different applications. Resource use measures that apply standardized prices allow for comparison of resource use units across regions and markets, while actual prices allow for comparison of prices paid which are often influenced by regional market conditions. The Committee found that an individual measure that allows both standardized and actual costing approaches has limited utility because differences in the measure score could be attributed to either to differences in resource use or differences in pricing and regional market conditions. Including both costing approaches within the same measure could reduce comparability and limit the user's ability to identify the source of variation.

Finally, measures submitted to this project spanned various levels of measurement analysis, from regional, to health plan, to individual provider. Measures specified at a higher level of measurement (i.e., health plan or regional) allowed for a comprehensive view of health service resource use by measuring all costs for a person across settings and providers. The burden of adjusting for comparability was lower for measures at the health plan level than it was for measures seeking to evaluate individual providers. When measures were specified at the individual provider level, and to a lesser extent at the group practice level, the Committee

engaged in a more detailed evaluation of the risk- adjustment approach and minimum sample size to ensure that the measures produced a reliable score.

Reporting

The reporting module includes steps for attribution, peer grouping, defining outliers and thresholds, sample size requirements, and benchmarking. These reporting steps could be submitted as measure specifications or guidelines, or could be left to the user's discretion. Specifications limit user options and flexibility and must be strictly adhered to, whereas guidelines are well thought-out guidance to users, which allow for user flexibility.

While sample size considerations could be submitted as guidelines or specifications in the reporting module, the Committee found that sample size was also relevant to the discussion of other modules and reliability and validity testing. In order to evaluate the number of patients required for a measure to demonstrate meaningful and statistically significant differences, the Committee encouraged measure developers to provide simulations and sensitivity analyses during the evaluation. When measures were specified at the individual provider level, confidence intervals need to be presented, especially when displaying information with small sample sizes. The use of confidence intervals allows the user to assess the estimated range of the measure score and true differences in provider performance.

Outliers were handled at both the episode and/or the claim level. During data preparation, high outlier claims were generally subject to a statistical technique used to limit the effect of extreme values and the effect of spurious outliers, known as *winsorization*.[11] Low cost claims were either winsorized or, more typically, were removed from measure analysis. Winsorization often sets outliers to a percentile of data; for example, all outliers above the 95th percentile are set to the value at the 95th percentile. Developers who chose to remove low-cost episodes indicated they took this approach because these episodes were likely to be incomplete and thus have the potential to skew the results. The Committee requested additional details from the developers on the effect of the winsorization and exclusion at the claim and episode-level on the measure score.

The experts noted that detailed listing and analysis of high-cost outliers could be useful for targeted improvement activities.

As part of the reporting module, the attribution approach could also be submitted as measure guidelines or specifications, or left to the user to define. The attribution approach is distinct from the level of analysis in that the level of analysis is the unit in which the measure has been tested and specified, while the attribution approach determines how the costs or resources are assigned to a provider, group of providers, health plan or region. Regardless of the approach submitted, the Committee agreed that it should reasonably allow for the accountable entity to affect the resource use of the patient. For example, if the attribution approach assigns a patient to the primary care provider (PCP) based on one evaluation and management (E/M) visit, the approach should not assign all of the previous hospitalization costs during the measurement year before the patient's first visit to this PCP. Proper consideration should be given to how the timing of patient encounters impacts the attribution rules and potential for unfair assignment of costs to providers. Lack of consideration for these types of factors creates the potential for unintended consequences of providers "gaming the system" to avoid attribution of extraneous costs to their profile for new patients with whom they have had limited contact.

Approach to disparities

Identifying and measuring disparities in care delivery is of critical importance to understanding variations in cost and improving quality. Gender and age were the most common factors accounted for in the stratification for disparities in the measures reviewed. The lack of information on race and ethnicity in commercial administrative data limited the ability of the resource use measures under evaluation to reflect disparities accurately in the results. Additional efforts should be pursued to capture this information more systematically. The Committee was unable to assess the measure's ability to identify disparities based on underlying limitations in the data. Measures were evaluated based on their ability to stratify if the underlying data included information on race and ethnicity.

Reliability and Validity testing

The next component to evaluating a measure's scientific acceptability is determining whether the measure testing approach and results demonstrate that the measure is reliable and valid. Reliability testing should demonstrate that the measure results 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. Validity testing must demonstrate that the measure data elements are correct and/or that the measure score correctly reflects the cost of care or resources provided, adequately distinguishing high and low resource use. If face validity is the only validity addressed, it must be assessed systematically. Reliability and validity testing can be demonstrated at the measure score or the data element level.

Data element reliability

Discussion of data element reliability was limited since resource use measures often relied on administrative claims data. Administrative claims provide accessible information on the processes of care and can generally be obtained as a byproduct of the care process. However, claims data provide only limited clinical information and lack granularity in determining patient health severity. Further, claims data are subject to variation in coding processes by the accountable entities. While these concerns are valid, the Committee agreed that they span measures of quality and resource use and are not limited the measures currently under evaluation.

Measure score reliability

Measure developers also performed varying levels of reliability assessments at the measure score level. Low levels of measure score reliability assessments depended on changes in measure specifications on the outcome variable (e.g., total resource use) to demonstrate measure score reliability. Higher levels of reliability assessments compared parallel development of episode grouper software and SAS using the exact same specifications. In some cases, reliability

demonstrated match rates of over more 99.9 percent. Developers whose measures have been in use were able to demonstrate the stability of the observed/expected results (O/E) over time.

The Committee suggested other robust methodologies that could be used to demonstrate a high level of reliability, including O/E ratio by accountable entities and conditions over time, to demonstrate score stability. These measures can also be tested using two independent programmers performing the same tasks to evaluate determine if the results are similar. Additional methods could include signal-to-noise ratio analysis using Analysis of Variance (ANOVA) or Intra-class Correlation Coefficient to demonstrate measure score reliability.

Data element validity

The validity testing submitted at the data element level was often weak, as there were no comparisons to other independent claims databases or other authoritative data sources. In addition to other claims databases, a comparison of the distribution of important variables to the literature would provide a more robust assessment of the validity of the data elements used.

With the exception of developers who require regular data audits to ensure data integrity, the measure submissions generally contained weak evidence of data integrity checks (i.e., percentage of missing values, missing diagnosis codes, or inconsistent dates). However, developers often provided guidelines for data preparation and missing data in the data protocol module.

Most measures submitted to the project were tested in large administrative claims databases representative of the target population. The Committee noted one exception in which a hip fracture measure was tested in a population with an age distribution outside of the age range in which the condition was most prevalent. The TAP agreed this testing approach calls to question the validity (and in fact the importance) of the measure as it has been tested and used to measure costs in a population where this condition is not high impact, and has limited clinical relevance.

Measure score validity

Validity testing at the measure score level often relied on face validity that the measure score was valid based on clinical review and empirical results. The measure score, however, was often not validated by correlating measure scores with other valid indicators, or by showing that the score produces different results when applied to subgroups known to have differences in resource use, as a more complex validity testing approach would demonstrate. Developers often demonstrated face validity by describing the distribution of measure score results, outlier status and type of service. While the Committee accepted this as a minimum threshold for demonstrating validity, they suggested more robust methods, including correlating the measure score with other valid indicators, should be applied in future iterations and testing.

Usability

The focus of the usability criteria is to determine whether the measure results are usable for the intended audience. This includes an evaluation of whether the measure is currently in use and the results are being reported for performance improvement and accountability purposes, and whether the results are considered meaningful and useful. For resource use measures, usability also includes the evaluation of whether it has been demonstrated that the measure construct and its components (e.g., risk-adjustment methodology, clinical logic) can be deconstructed to enable transparency and understanding.

Resource use measures presented some specific challenges to applying the concepts identified within the usability criterion. For example, the issue of accountability is a charged one. No consensus existed as to who the intended audience of these measures should be: purchasers, the public at large (consumers), health plans, and health plan members, are all likely users of this information. It was noted that for the public at large, extra effort would be required to make the reporting of these measure results as clear as possible. This clarity is the focus of consumer-oriented organizations that share data such as these. There was agreement that these measures should not be reported alone, but in the context of quality measures.

Another challenge the TAPs and Committees encountered was differentiating between usability and usefulness and determining whether a measure is inherently usable because it is in use. For measures not currently in use, they questioned how usefulness should be demonstrated since there is a lack of knowledge of the practical application of the measure.

The Committee also questioned the usability of measures that are embedded in a complex episode-grouper system in which each individual measure's logic is interwoven and tied to the logic of another measure, which may not be under evaluation. They struggled with how to evaluate the usability of a single measure without evaluating the entire grouper system.

The final overarching issue identified within the usability criteria relates to transparency. Many of the TAP and Committee members expressed concern over the complexity of certain methodologies used and questioned whether this complexity masks these measure's ability to be transparent. Difficulty understanding how the risk-adjustment, severity level assignments, and episode logic work together in a measure may make it difficult for a physician, for example, to completely understand completely which of his or her patients have been included in the costs attributed to them and how the complexity of the patient population has been accounted for in those costs. Some Committee members argued that this lack of transparency and understanding of the construction logic affects the ability of the reported measure score to be used and may limit the physician or health plan from identifying how and where to improve scores. Committee members also questioned whether there should be an expectation that these complex measures would require an investment of time to be interpreted and understood. It was pointed out, however, that by using the resource use service categories identified within the measure, action could be taken using the categories in which high costs were most evident (e.g., imaging, outpatient visits).

Feasibility

The feasibility criterion focuses on the extent to which the measure can be implemented with undue burden and identifies any barriers to implementation. The feasibility subcriteria used to evaluate the resource use measures are identical to those used to evaluate quality measures.

Because all of the resource use measures submitted to this project solely rely solely on the use of administrative claims data, the subcriteria evaluating the availability of required data via electronic sources and whether the data are routinely generated required very little discussion. The remaining feasibility subcriteria, however, illuminated some important issues related to the implementing of resource use measures, which often use very complex, sophisticated methodologies to risk adjust and determine episode logic, for example. This issue of complexity for the implementer (and for the users of the results) was discussed at length by the TAPs and the Committee during their evaluation of susceptibility to errors and inaccuracies. Some members expressed concern that the complexity of the methodologies lends itself to user error, most likely on behalf of the programmer who would develop the code to run the measures. This issue may be mitigated by the purchase of a product that is pre- programmed to implement the measure with imported data or the submission of data to an organization that audits, computes the measure, and reports the information back to the user.

Additionally, having been in use in the market place by health plans and purchasers for many years, these measures often use some proprietary component or are imbedded in sophisticated proprietary products. For product lines that include large episode-grouping tools encompassing many conditions, a user would be required to purchase some or parts of a product suite to run a single episode for diabetes, for example. For this reason, the feasibility of implementing an individual clinical episode may be very limited. The Committee expressed concern that the financial burden on a practice or system to purchase these products could be very significant, thus creating a barrier to measuring resource use applying NQF-endorsed standards.

Similar Measures

In phase one of this resource use measurement project, the Committee agreed that since this is NQF's first effort focused on evaluating resource use measures, identifying "best-in-class" and requiring harmonization among resource use measures was premature. While the Committee would forgo the selection of "best-in-class" measures, they would discuss the merits of and justify the recommendation for similar measures and discuss potential ways in which

harmonization among related and similar measures might be achieved. In the context of resource use measures, similar measures are defined as the same measure types (e.g., per episode, per capita) measuring the same costs/resources (e.g., actual cost vs. standard prices, resource service categories) in the same population (e.g., patients with diabetes). The Committee will discuss all related and similar measures and potential for harmonization of resource use measures at the conclusion of the cycle two review process once all endorsement recommendations are complete.

RECOMMENDATIONS FOR ENDORSEMENT

This report presents the results of the evaluation of four measures considered under NQF's CDP.

Candidate Consensus Standards Recommended for Endorsement

Four measures are recommended for endorsement as voluntary consensus standards suitable for accountability and performance improvement.

The evaluation summary tables follow the list of measures and summarize the results of the TAP's and Committee's evaluation of and voting on the candidate consensus standards that were recommended for endorsement. Hyperlinks are provided from each summary table to the detailed measure specifications. To access the meeting transcripts and recordings in which these measures are discussed, refer to the project web page.

The Committee recommended the following candidate consensus standards for endorsement:

Evaluation Summary—Candidate Consensus Standards Recommended for Endorsement

1557: Relative Resource Use for People with Diabetes
Description: The risk-adjusted relative resource use by health plan members 18-75 years of age who were identified as having diabetes (type 1 and type 2) during the measurement year. Resource Use Measure Type: Per capita (population- or patient-based) Data Source: Administrative claims Resource Use Service Category: Inpatient services: Inpatient facility services; Inpatient services: Evaluation and management; Inpatient services: Procedures and surgeries; Inpatient services: Imaging and diagnostic; Inpatient services: Lab services; Inpatient services: Admissions/discharges; Ambulatory services: Outpatient facility services; Ambulatory services: Emergency Department; Ambulatory services: Pharmacy; Ambulatory services: Evaluation and management; Ambulatory services: Procedures and surgeries; Ambulatory services: Imaging and diagnostic; Ambulatory services: Lab services Care Setting: Ambulatory Care: Clinic/Urgent Care; Ambulatory Care: Clinician Office; Hospital/Acute Care Facility; Imaging Facility; Laboratory; Pharmacy Level of Analysis: Health Plan; Integrated Delivery System; Population: National; Population: Regional Measure Developer: National Committee for Quality Assurance (NCQA)
Committee Recommendation for Endorsement: Y-17; N-0; Abstain-1
If applicable, Conditions/Questions for Developer and Developer response: • In relation to criterion 2a.1, provide information on which maternity codes are included. • In relation to criterion 2b.3, provide rationale for excluding patients >75 years old.
TAP Evaluation: 1. Importance to Measure and Report: 1a.High Impact: H- 9 , M-0, L-0, I-0, N/A-0 *TAP Discussion:* Developer provided sufficient evidence and support. 1b. Resource use/cost problems: H- 9, M-0, L-0, I-0, N/A-0 *TAP Discussion:* Developer provided sufficient evidence and support. 1c. Purpose clearly described: H- 8, M-1, L-0, I-0, N/A-0 *TAP Discussion:* Developer provided sufficient evidence and support. 1d. Resource use service categories consistent and representative: H- 7, M-2, L-0, I-0, N/A-0 *TAP Discussion:* The resource use service categories were sufficient.
Overall Importance: Y-17, N-0 Committee Discussion: While the measure is deemed important, the Committee pointed out the resources accounted for in the measure do not include important services provided to diabetic patients, including care coordination, and education by nurses and nutritionists. These services are typically not billed, services and so they are often left out of the cost calculations for measures using administrative claims data. This type of measurement is possible, but NCQA does not generally have access to this level of specificity in the data only at the utilization level.
TAP Evaluation 2. Scientific Acceptability of Measure Properties: 2a1. Well defined/precise specifications: H- 8, M-0, L-0, I-0 *TAP Discussion:* The TAP had concerns about how are changing codes are handled. It was stated that this is very difficult to manage in all measures. Concern was also expressed related to adjusting away patients with lots of claims; conditions such as HIV and active cancer are excluded (this adjustment is made every year with a one year lag). The intent of this measure is to capture all costs for a diabetic patient, including services that may not be related to a diabetes diagnosis.

1557: Relative Resource Use for People with Diabetes
While counting all costs does add some noise to the measure, there is evidence that diabetics stay in hospital longer, even for stays triggered by non-diabetes related events. With a minimum sample size of 400, this measure has been specified for use at the health plan level; not for use at the physician attribution level. TAP had concerns as to why conditions that are proven to be related to diabetes complications are not included, for example, amputations, ESRD, etc. The TAP wanted clarification on whether pregnancy/maternity codes were included in this measure. 2a2. Reliability testing: H- 9, M-0, L-0, I-0, N/A-0 *TAP Discussion:* Reliability testing was acceptable. 2b1. Specifications consistent with resource use/cost problem: H- 5, M-4, L-0, I-0, N/A-0 *TAP Discussion:* Measure captures all costs for a diabetes patient. 2b2. Validity testing: H- 5, M-4, L-0, I-0, N/A-0 *TAP Discussion:* Adequate validity testing information provided. 2b3. Exclusions: H- 6, M-3, L-0 ,I-0, N/A-0 *TAP Discussion:* The TAP expressed concern over the age limit criteria; Age 75 may be too low. 2b4. Risk-adjustment : H- 9, M-0, L-0, I-0, N/A-0 *TAP Discussion:* Measure uses HCC's for the risk-adjustment. The TAP agrees this is acceptable methodology. 2b5. Identification of statistically significant/meaningful differences: H- 9 , M-0, L-0 , I-0,N/A-0 *TAP Discussion:* Minimum sample size at 400 allows for increased statistical stability. 2b6. Multiple data sources: H- 0, M-0, L-0 , I-0, N-9, N/A-0 *TAP Discussion:* N/A 2c. Stratification for disparities: H- 2, M-5 , L-1, I-0,N-, N/A-0 *TAP Discussion:* Can only be stratified only for age, gender and region, as with most of the measures submitted.
Overall Scientifically Acceptable: Yes [Y-18; N-0 (Committee Vote)] Committee Discussion: There was acknowledgement that certain types of claims and clinicians are invisible in these types of measures because administrative claims data does not capture all resource use or recognize the resources used by of all types of clinicians. The Committee also pointed out that a broad scope of cost codes are going to be important, and the thinking about measuring resources should be expanded beyond intermediate care and consider home health costs, skilled nursing facilities, etc. There was discussion on the use of the standardized pricing tables and how they are applied within the measures. These pricing tables are now publicly available on the NCQA website and can be used by anyone for their own purposes. A number of resources have been used to develop the tables, including the Medicare fee schedule and data from thousands of pharmacy prescriptions. The TAP identified concern over the exclusion of patients over the age of 75 identified by the TAP. The TAP also identified concern over the mandatory exclusions for active cancer, transplantation, ESRD, and HIV that are applied to all NCQA measures, but are particularly relevant to the diabetes population. The developers are going back to re-examine these exclusions for future versions of the measure. The final concern the Committee addressed related to the logic of truncation scheme. In order to avoid a small proportion of members driving up the standardized costs, the developers identified cap levels at which members would be capped and truncated once costs reach that high level; however, they are not excluded. This also prevents skewing of the results. The timeframes used in the measure logic were in attempt to focus on a group of patients who are not newly diagnosed.
TAP Evaluation: 3. Usability: 3a. Measure performance results are publicly reported: H- 9, M-0 , L-0 , I-0, N/A-0 *TAP Discussion:* Measure is currently in use by large number of health plans. 3b. Measure results are meaningful/useful for accountability and quality improvement: H- 8, M-1 ,L-0 , I-0, N/A-0 *TAP Discussion:* Accountability mechanism sufficient. 3c. Data and results can be decomposed for transparency and understanding: H- 8, M-1, L-0, I-0, N/A-0 *TAP Discussion:* Specifications adequate for transparency. 3d. Harmonized or justification for differences: N/A *TAP Discussion:* Developers were not asked to harmonize prior to submissions. Harmonization may come up as the set of measures move through the CDP process.
Overall Usability: H-12; M-6; L-0; I-0 Committee Discussion: The Committee did not identify any additional issues for this criterion.

1557: Relative Resource Use for People with Diabetes
TAP Evaluation: 4. Feasibility: 4a. Data elements routinely generated during care process: H- 9, M-0, L-0 , I-0, N/A-0 *TAP Discussion:* Measures rely on administrative data. 4b. Data elements available electronically: H- 9, M-0, L-0 , I-0, N/A-0 *TAP Discussion:* Administrative data are in electronic format. 4c. Susceptibility to inaccuracies/ unintended consequences identified: H- 6,M-3, L-0, I-0, N/A-0 *TAP Discussion:* Users of NCQA are subject to a data audit process. Susceptibility to errors/inaccuracies is low. 4d. Data collection strategy can be implemented: H- 9, M-0, L-0 , I-0, N/A-0 *TAP Discussion:* Barriers to use are low.
Overall Feasibility: H-11; M-7; L-0; I-0 Committee Discussion: There were no additional concerns identified by the Committee.

703

1558 Relative Resource Use for People with Cardiovascular Conditions
Description: The risk-adjusted relative resource use by health plan members with specific cardiovascular conditions during the measurement year. Resource Use Type: Per capita (population- or patient-based) Data Type: Administrative claims; Electronic Clinical Data: Electronic Health Record; Electronic Clinical Data: Imaging/Diagnostic Study; Electronic Clinical Data: Laboratory; Electronic Clinical Data: Pharmacy; Paper Records Resource Use Service Category: Inpatient services: Inpatient facility services; Inpatient services: Evaluation and management; Inpatient services: Procedures and surgeries; Inpatient services: Lab services; Inpatient services: Admissions/discharges; Ambulatory services: Outpatient facility services; Ambulatory services: Emergency Department; Ambulatory services: Pharmacy; Ambulatory services: Evaluation and management; Ambulatory services: Procedures and surgeries; Ambulatory services: Imaging and diagnostic; Ambulatory services: Lab services Care Setting: Administrative claims, Cardiovascular: Cardiovascular; Electronic Clinical Data : Electronic Health Record; Electronic Clinical Data: Imaging/Diagnostic Study; Population Health: Population Health Level of Analysis: Administrative claims: Administrative claims, Cardiovascular; Electronic Clinical Data : Electronic Health Record; Electronic Clinical Data: Imaging/Diagnostic Study, Population Health Measure Developer: National Committee for Quality Assurance (NCQA), 1100 13th Street NW, STE 1000, Washington, DC, 20005
Committee Recommendation for Endorsement: Y-13; N-3; Abstain-1
TAP Conditions/Questions for Developer: 1. Are other conditions similar to Coronary Artery Disease included, such as ischemic heart disease? 2. How does the stratification discern between high- and low -risk patients? 3. What is the time frame for exclusions? 4. How would a provider know how to improve based on the report? Developer Response: 1. This measure is based on the HEDIS measure, covering both acute and sub-acute, ischemic heart disease, cardiovascular unspecified, angina, atherosclerosis of extremity, etc. CAD-related codes diverged into family history, etc. The measure does not try to account for anything other than what CAD is described as in the code set. The developer is going to look into including code sets that are non -CAD -specific for non-traditional patients. 2. In terms of stratification for the risk-adjustment, it is dependent on the number of comorbidities. Section 10.1 includes additional information on the risk -adjustment methods, identifies based on qualifying and HCC rankings. 3. The time frames align with the eligible population period; patients are looked at a year prior to the measurement year and are looked at the year prior to and during the eligibility period. 4. The reports are divided up by resource categories; user would need to look into measure specifications, which are fairly broad. Committee Follow-up: • Has this type of risk -adjustment model been validated in the past? HCC are well validated. RTI evaluated this in April 2011, and it continues to be a valid stratification method. • The Committee wanted additional follow-up on the time period for eligibility for risk-adjustment/ exclusions.

If Applicable, Questions to the Committee: N/A

TAP Evaluation:
1.Importance to Measure and Report
1a.High Impact: H-5; M-0; L-0; I-0
TAP Discussion: The TAP agreed that this subcriterion has been met.
1b. Resource use/cost problems: H-5; M-0; L-0; I-0
TAP Discussion: The TAP agreed this subcriterion has been met and is supported by the evidence.
1c. Purpose clearly described: H-3; M-3; L-0; I-0
TAP Discussion: Inclusion criteria for this measure are very broad – PCI and CABG, but not other codes are associated with chronic conditions. It would be difficult for this measure to be actionable by an individual provider because of the broad nature of the category. The costs of carotid disease are included in the category. It does capture costs, but there is the issue of which costs are incorporated and which costs are not. Given the broad category, the calculation of costs is difficult for a user to understand. This measure covers all costs across all procedures and excludes those who were screened and had plaque in their carotid paired equally as with those with PCIs and that early detection may become a preponderance of those grouped in cardiovascular disease.
1d. Resource use service categories consistent and representative: H-2; M-3; L-0; I-0
TAP Discussion: The TAP agreed that this subcriterion has been met.

Overall Importance: Y-14; N-1; Abstain-1
Committee Discussion: There were no additional concerns identified by the Committee for this criterion.

TAP Evaluation:
2.Scientific Acceptability of Measure Properties:
2a. Reliability:
2a1. Well- defined/precise specifications: H-2; M-1; L-1, I-0
TAP Discussion: The specifications don't consider the cost; rather, they use what RVUs would be, i.e., the actual resource use versus the cost. The Committee believed this to be a relevant way to approach the measure, as each grouping and person is stratified according to risk. It is unclear which risk-adjustment is used for each patient. This measure is calculated by using databases from insurers, up to age 75, and only reports only on organizations with more than 400 people in the measure. This measure is restricted in use for larger groups.
2a2. Reliability testing: H-2; M-2; L-0; I-0
TAP Discussion: The reliability testing uses data from 15 months. The results are consistent with other models.
2b. Validity:
2b1. Specifications consistent with resource use/cost problem: H-1; M-2; L-2; I-0
TAP Discussion: Discussion similar to 2a1. It is unclear which risk-adjustment is used for which patient.
2b2. Validity testing: H-2; M-2; L-0; I-0
TAP Discussion: NCQA publicly reported the results annually and continues to publicly report publicly. The costs are standardized and are good measures of the resources being used. There is a track record of data being clean, including resource use not what was actually charged.
2b3. Exclusions: H-1; M-2; L-1; I-1
TAP Discussion: The measure is unclear regarding the time period for exclusions.
2b4. Risk-adjustment: H-1; M-2; L-1; I-0
TAP Discussion: It is difficult to discern what is included in risk -adjustment criteria. Unclear how stratification is working and if the groups produced is are legitimate.
2b5. Identification of statistically significant/meaningful differences:
TAP Discussion: The Committee has agreed this subcriterion has been met.
2b6. Multiple data sources: H-1; M-4; L-0; I-0; N/A-0
TAP Discussion: N/A
2c. Stratification for disparities: H-0; M-4; L-0; I-0; N/A-1
TAP Discussion: This measure stratifies for age and gender.

Overall Scientifically Acceptable : Yes [Y-13; N-4 (Committee Vote)]
Committee Discussion: Submission form level of analysis check boxes need to be fixed to show only health plan level. Concerns with comparing like plans (e.g., Medicaid to Medicaid plans). The measure submitted must be used at health plan level, as the Committee was very uncomfortable with using measure at physician or group level. Developer acknowledged that there has been testing of the measure at the group practice level; however, it was only tested with over 400 patients. The Committee was interested in the exclusions

for end stage renal disease (ESRD). The Committee was concerned with the peer group comparison of "like plans" because there might be correlations with socioeconomic status (SES) across plans. Further, the Committee was concerned over the appropriateness of excluding patients who are >75 years old.
The risk -adjustment model used in this measure includes HCCs where risk-adjustment takes into account the resource use from within the measurement year. The Committee agreed that a better title for the measure might be "Measure of Patients with Chronic Cardiac Conditions."
While the Committee was concerned with the level of measurement, the developer clarified that it would only be used at population level, and reported with quality measures. Purchasers and health plan representatives agreed that this measure would be useful.

TAP Evaluation:
3. Usability:
3a. Measure performance results are publicly reported: H-3; M-1; L-0; I-0
TAP Discussion: This measure has been utilized for a short amount of time (since 2007); it is difficult to assess if the manner in which they are reporting is useful.
3b. Measure results are meaningful/useful for accountability and quality improvement: H-2; M-1; L-1; I-0
TAP Discussion: There is no data on how consumers are utilizing the data and making changes based on this measure. It is unclear what would or would not affect the score and change practices in the long run. The measure would rate fairly low for this subcriterion. It may not be extremely useful for accountability as it's it is not easily interpreted.
3c. Data and results can be decomposed for transparency and understanding: H-2; M-1; L-1; I-0
TAP Discussion: The measure is very broad and it's unclear how providers can change behavior.
3d. Harmonized or justification for differences: N/A
TAP Discussion: N/A

Overall Usability: H-6; M-9; L-2; I-0
Committee Discussion: TAP was comfortable with the measure since it has been in use for 5 years (with focus groups). It expressed concern over how the results will be used for consumers. The breakdown within the service categories was found to be more useful information than the overall score. There are currently 800 out of 1100 plans reporting RRU/quality measures with less than 1% of the health plans as outliers.
The Committee was not as concerned with "carve -outs" since pharmacy costs are reported separately from medical costs. There was interest in how to make this kind of data could be meaningful for consumers as well. The developer clarified that the major users are employer groups and business groups, and it helps to inform their decisions for the following year. However, skepticism had been expressed regarding the usability at the plan level.

TAP Evaluation
4. Feasibility:
4a. Data elements routinely generated during care process: H-4; M-0; L-0; I-0
TAP Discussion: All administrative data is generated as a byproduct of care.
4b. Data elements available electronically: H-4; M-0; L-0; I-0
TAP Discussion: All data is available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-1; M-4; L-0; I-0
TAP Discussion: This subcriterion has been met.
4d. Data collection strategy can be implemented: H-4; M-0; L-0; I-0
TAP Discussion: This subcriterion has been met.

Overall Feasibility: H-7; M-6; L-3; I-1
Committee Discussion:
The developer explained that health plans calculate observed measure scores but NCQA calculates the expected for the final measure score. The Committee was interested in how carve-outs and capitated arrangements were addressed. Data within the measure is are reported out into each service categories with pharmacy benefits measured separately.

704

705

1598 Total Resource Use Population-based PMPM Index

Description: Total cost of care reflects a mix of complicated factors such as patient illness burden, service utilization, and negotiated prices. Separating out and reporting the resource use index along with the total cost of care index provides a more complete picture of population- based drivers of health care costs. Total Cost Index (TCI) is a measure of a primary care provider's risk-adjusted cost

effectiveness at managing the population for which they care for. TCI includes all costs associated with treating members, including professional, facility inpatient and outpatient, pharmacy, lab, radiology, ancillary, and behavioral health services. The Resource Use Index (RUI) is an underlying risk-adjusted measure of the frequency and intensity of services utilized to manage a provider group's patients. Resource use includes all resources associated with treating members, including professional, facility inpatient and outpatient, pharmacy, lab, radiology, ancillary, and behavioral health services.
Resource Use Type: Per capita (population- or patient-based)
Data Type: Administrative claims, other
Resource Use Service Category: Inpatient services: Inpatient facility services; Inpatient services: Evaluation and management; Inpatient services: Procedures and surgeries; Inpatient services: Imaging and diagnostic; Inpatient services: Lab services; Inpatient services: Admissions/discharges; Inpatient services: Labor (hours, FTE, etc.); Ambulatory services: Outpatient facility services; Ambulatory services: Emergency Department; Ambulatory services: Pharmacy; Ambulatory services: Evaluation and management; Ambulatory services: Procedures and surgeries; Ambulatory services: Imaging and diagnostic; Ambulatory services: Lab services; Ambulatory services: Labor (hours, FTE, etc.); Durable Medical Equipment (DME)
Care Setting: Ambulatory Care: Ambulatory Surgery Center (ASC); Ambulatory Care: Clinic/Urgent Care; Ambulatory Care: Clinician Office; Behavioral Health/Psychiatric: Inpatient; Behavioral Health/Psychiatric: Outpatient; Dialysis Facility
Level of Analysis: Clinician: Group/Practice; Population: Community
Measure Developer: HealthPartners, 8170 33rd Avenue South, PO Box 1309, Bloomington, Minnesota, 55425

Committee Recommendation for Endorsement: Recommended for Endorsement: Y-11 ; N-6

Committee Questions for Developer:

1. The measure's resource use index relies on total care relative resource use categories, which are constructed so they are additive across various sites of care and then add in pharmacy data. How was this done?
2. Are the data distorted due to billed charges?
3. What is the attributable population in this measure?
4. How are variables in geographic location accounted for?
5. This measure is restricted to commercial, under -65 -years -of -age population. Is there anything that prohibits its use in the Medicare population?
6. Do users have to use the ACG software for risk-adjustment?

Developer Responses:

1. Health Partners relies on sector- specific relative value units, the billed charges across the sectors of care are used to build relativity. The payments are then appropriately adjusted. Final quality checks for thresholds are then performed. This method will eventually be patented and shared with the community.
2. The measure uses billed charges controls for confounding variables. The measure uses the billed amount to allow for the claims (the most standard piece of information), then goes across the different components and applies the discount rate. The adjustment factor is for the paid/billed ratios.
3. The attributable populations (which are scalable to different units of analysis) are PPO and HMO. Look at practice specialty of physician and claims history and attribute patients to the clinic with the majority of visits.
4. Depending on the application and the user, the measure can be flexible and usable across different locations. In the market there are multiple hospitals with different price points. Cost points may be consistent; however, the price they charge may be different. Actual paid (allowable inclusive liability) amount is used in the measure; the billed amount is used only to gauge the relativity (e.g., inpatient to outpatient services).
5. HealthPartners is a largely commercial- based health plan, so they do not have access to Medicare data. Theoretically, if these claims were available in the database, one would be able to use it.
6. Users are not required to use the ACG software for risk-adjustment. Any risk- adjustment methodology may be used, as long as all methods are comparable (see Society of Actuaries report). Health Partners has a history of working with ACG software and have tested the measure using the ACG risk adjuster. They have specified the measure to be used at the group level with the risk- adjustment methodology developed by Johns Hopkins, and if it is NQF-endorsed, it would be endorsed only at the group level for use with this specific software.

Committee Conditions:

1. The Committee determined there were actually two measures of cost described within the measure submission as presented: resource use index and a total cost index. There was some discussion about which should be evaluated for the purposes of this project or whether the measures should be considered as a pair. Because this project is not accepting paired measures, the Committee has agreed to evaluate the resource use index, which appears to be most

applicable to the goals of this project at this time.

Developer Response:

1. The measure calculations for costing within the measure may be used independently; however, they are better used in partnership with one another. The developer agreed to separate the specifications and resubmit a separate measure for total cost (#1604).

**Please note: NQF endorses the measures only for the populations in which it was tested.*

1.Importance to Measure and Report

1a.High Impact: H-15, M-2, L-1, I-0, N/A-0

Committee Discussion: This measure is considered highly important and relates to NPP/national goals.

1b. Resource use/cost problems: H-13, M-3, L-0, I-1, N/A-1

Committee Discussion: This measure does not explain much as an isolated measure. However, it does inform providers of areas where there is overuse or underuse; given the fact of that overuse and waste is an issue, there is a place for this in the resource use project.

1c. Purpose clearly described: H-12, M-5, L-1, I-0, N/A-0

Committee Discussion: This criterion has been met because the measure is targeting an area known to have variation, and relevant service categories, and the objective has been clearly described.

1d. Resource use service categories consistent and representative: H-12, M-6, L-0, I-0, N/A-0

Committee Discussion: This criterion has been met. The supporting information provided by the measure developer also helps to demonstrate this.

1.Overall Importance: Y- 18, N-0

2.Scientific Acceptability of Measure Properties:

2a. Reliability:

2a1. Well- defined/precise specifications: H-5, M-8, L-4, I-0, N/A-0

Committee Discussion: HealthPartners (HP) uses regional and national data; there is a great deal of actionable data at this level. It may be difficult to be implemented in other systems. Since this is a population measure, it is missing whether or not people are described on an individual basis and then tied to a region, making it difficult to determine whether or not it was appropriately specified. The total eligible individuals may only have pharmacy claims or are not using any services; however, this may vary across systems. This measure is intended for a commercial population; non-users would not be attributed. The patient has to be a user of primary care services to be included; attribution (prospective and retrospective) is at the physician group level (with 2 or more physicians). The peer groups are based on the group to which the physician belongs to. The measure has been tested on groups that have at least 600 patients at the group practice level. High claims data are included and truncated after a certain threshold, resulting in roughly 5-8% excluded. These individuals are excluded based on the published guidelines by Society of Actuaries. The pharmacy relative values come from using the average billed amount, and the paid amount is defined as the paid-to-billed ratio.

2a2. Reliability testing: H-10, M-6, L-0, I-1, N/A-0

Committee Discussion: Assumption that clinical and administrative claims data is accurate from a coding perspective, which is true for the majority of resource measures. For claims data, the hospital-based claims take more time to process than professional claims, so time frames need to be taken into account when applying them to this measure. The measure developer informed the Committee that the timeline of 3 months is specified; all claims are electronic and therefore arrive quickly into the system. The Committee believes the reliability matrix is acceptable. Health Partners did a very good job examining the reliability of the data using its commercial database. They performed two types of sampling; the first was a 90% sample of the actual values. It selected one patient at a time until they it reached 90%; this gives an idea of the influence of extreme values. Health Partners selected 90% of the data 500 times and compared the results obtained from the averages to the entire sample; the results showed there is represent very small change. The difference between the samples is only 0.9%, so that demonstrating reliability and that the potential influence of these extreme values is small. The other approach used was a boot strapping technique, which is similar; but instead of a 90% sample, however, the developers selected a sample with replacement, this simulates the reliability and is a very common methodology. The developers found a very small range of change in the sample population; this has some variability in respect to the sample. It's important to note that NQF does not require developers use a certain type of methodology. The analysis has been done at the provider level and depicts the measure to be reliable.

2b. Validity:

2b1. Specifications consistent with resource use/cost problem:

Committee Discussion: This section appeared to be sufficient and meets the criterion. This measure excludes patients who have not had a primary care visit; however, within the system this may be giving all the information needed to feed back to providers on how they are using services.

2b2. Validity testing: H-5, M-8, L-2, I-1, N/A-0
Committee Discussion: Adequate sample size, large area, 19 providers across approximately 200 hospitals. Health Partners (HP) has nearly 7,000 members who are Medicare/Medicaid recipients. HP has about 700,000 total members within the marketplace area (including CMS data/commercial data), and the non-user rate is around 9%. Roughly 50% of the data presented in the validity sample comes from commercial data. Because this measure has only been tested only in a commercial population, it will be NQF endorsed only in a commercial population. Peer group averaging can serve as a benchmark, if that is a sufficient measure in all markets. Within a commercial network and scheme, it may work; however, how these will be used it is not clear how these will be utilized. The validity was obtained in terms of the risk -adjusted and the non-risk adjusted values. One would anticipate the values between expected and observed would be close - values of 0.98 for non-risk adjusted to actual money spent. After the measure risk-adjustment was applied, this correlation went down to 0.215. When the correlation is restricted to different places, they look at the correlation between total resource use to the risk adjusted methods. There were a number of test performed and they show the direction of the correlation, which was high in this case.
2b3. Exclusions: H-6, M-8, L-1, I-2, N/A-0
Committee Discussion: This measure excludes sub-populations that haven't had primary care visits. The measure also excludes "never users" and "super users" by truncating them out. The group-oriented market may exclude those outside the group. HP has not seen this as a problem, as there is a low non-user rate. The bulk of members are attributed in this model through primary care, a smaller percentage only see a specialist. Those who are over the age of 65 are excluded.
2b4. Risk-adjustment: H-5, M-9, L-0, I-3, N/A-0
Committee Discussion: Health Partners uses the 9.0 version of the ACG risk-adjustment method developed by Johns Hopkins, the most recent 9.0 version and they HP has a long-standing market history of using this product. HP relied heavily on a study conducted by the Society of Actuaries that concluded a number of commercially available risk-adjustment methodologies are satisfactory for this purpose. The risk-adjustment was tested and demonstrated to be effective. It is significant to note for consumers that a user ID and password is necessary to access the site. The Johns Hopkins software is proprietary; however, Hopkins has recently announced the software to be free of charge to health insurance exchanges. For the ordinary user, the software is available for a fee based on a scale from large to small organizations, non-profits, etc. CMS offers an open -source risk -adjustment tool, the Hierarchical Condition Categories (HCCs).
2b5. Identification of statistically significant/meaningful differences: H-7, M-6, L-1, I-3, N/A-0
Committee Discussion: The Committee believes that this sub criterion has been met.
2b6. Multiple data sources: N/A
Committee Discussion: N/A
2c. Stratification for disparities: N/A
Committee Discussion: N/A

2.Overall Scientifically Acceptable: Yes [Y-13, N-4 (Committee Vote)]

3. Usability:
3a. Measure performance results are publicly reported: H-5, M-5, L-2, I-2, N/A-2
Committee Discussion: The data is publicly reported, but it's difficult to find on the Health Partners website. Currently the measure is used for benefit design and transparency; there are plans for community collaborations in the future.
3b. Measure results are meaningful/useful for accountability and quality improvement: H-3, M-9, L-3, I-1, N/A-2
Committee Discussion: The Committee discussed the issue that publicly reported measures may not have the same value for quality improvement. This measure is being reported out to the public at large, as well as to members of Health Partners, and has been for quite some time. During the three-year NQF maintenance review this criterion would be looked at even further to see how the measure has progressed. This is a fairly complicated measure for the public, in that the methodology may not be fully understandable to the average person. It must be communicated that more resource use does not necessarily mean *better* service. For resource use, it may be up to those producing the consumer reports on may need to be the ones determining how to present it to the public in the most understandable way.
3c. Data and results can be decomposed for transparency and understanding: H-3, M-8, L-4, I-2
Committee Discussion: On the Health Partners website, they have converted the results to dollar signs. This calculation is available to the public at large. There have also been focus groups conducted in order to gauge the clarity of the information available online. It may be difficult to decipher differences in providers and resource use; at some point there is the issue of hierarchical modeling and how to devise low -volume providers by evaluating the measure itself occurs. To some extent, the issue is raised are whether the measure is useful to the public because it does not explain the quality of care or outcome relating to resource use.
3d. Harmonized or justification for differences: N/A

Committee Discussion: N/A
3.Overall Usability: H-3, M-11, L-2, I-1, N/A-0
4. Feasibility: 4a. Data elements routinely generated during care process: H-9, M-1, L-3, I-0, N/A-4 *Committee Discussion*: This measure is based on data that is generated as a byproduct of care. The Committee believes this criterion has been met. 4b. Data elements available electronically: H-13, M-0, L-1, I-3, N/A-0 *Committee Discussion*: These measures are all available via electronic sources. The Committee believes this criterion has been met. 4c. Susceptibility to inaccuracies/ unintended consequences identified: H-4, M-8, L-3, I-2, N/A-0 *Committee Discussion*: This measure has met the criteria for inaccuracies and unintended consequences. Third- party administrators can work together to match up their coding; this would not be a barrier for these measures. There is a great deal of regulatory variation that can be applied to self-insured entities, and runs the risk of measuring smaller percentages of practices. 4d. Data collection strategy can be implemented: H-5, M-5, L-4, I-2, N/A-0 *Committee Discussion*: The Committee believes this sub criterion has been met.
4. Overall Feasibility: H-7, M-7, L-1, I-1, N/A-0

706

1604 Total Cost of Care Population-Based PMPM Index
Description: Total Cost of Care reflects a mix of complicated factors such as patient illness burden, service utilization, and negotiated prices. Total Cost Index (TCI) is a measure of a primary care provider's risk-adjusted cost effectiveness at managing the population they care for. TCI includes all costs associated with treating members, including professional, facility inpatient and outpatient, pharmacy, lab, radiology, ancillary, and behavioral health services. A Total Cost of Care Index when viewed together with a resource use measure provides a more complete picture of population-based drivers of healthcare costs. Resource Use Type: Cost/resource use Data Type: Administrative claims Resource Use Service Category: Inpatient services: Inpatient facility services; Inpatient services: Evaluation and management; Inpatient services: Procedures and surgeries; Inpatient services: Imaging and diagnostic; Inpatient services: Lab services; Inpatient services: Admissions/discharges; Inpatient services: Labor (hours, FTE, etc.); Ambulatory services: Outpatient facility services; Ambulatory services: Emergency Department; Ambulatory services: Pharmacy; Ambulatory services: Evaluation and management; Ambulatory services: Procedures and surgeries; Ambulatory services: Imaging and diagnostic; Ambulatory services: Lab services; Ambulatory services: Labor (hours, FTE, etc.); Durable Medical Equipment (DME) Care Setting: Ambulatory Care: Ambulatory Surgery Center (ASC); Ambulatory Care: Clinic/Urgent Care; Ambulatory Care: Clinician Office; Behavioral Health/Psychiatric: Inpatient; Behavioral Health/Psychiatric: Outpatient; Dialysis Facility; Emergency Medical Services/Ambulance; Home Health; Hospice; Hospital/Acute Care Facility; Imaging Facility; Laboratory; Pharmacy; Post-Acute/Long Term Care Facility: Nursing Home/Skilled Nursing Facility; Post-Acute/Long Term Care Facility: Rehabilitation Level of Analysis: Clinician: Group/Practice; Population: Community Measure Developer: HealthPartners, 8170 33rd Avenue South, PO Box 1309, Bloomington, Minnesota, 55425
Committee Recommendation for Endorsement: Recommended for Endorsement: Y-9 ; N-8, Abstain- 1
Committee Conditions/Questions for Developer: 1. What tools are used to collect patient satisfaction information? 2. In this measure it appears the total cost measure is reduced to an index and then compared to a peer group. Is it correct that any variations in input costs should be factored into that peer group comparison? 3. How are regional comparisons made between regions with very different cost/payment structures? 4. Are the actual prices based on what the plan has paid or what has been billed? 5. Have you tested this measure within a system that uses behavioral or pharmacy carve-outs? 6. When the costs per member per month (PMPM) are calculated, is this the average premium they are paying for the carve-out for every member in the group specific, or is it adjusted to reflect it? 7. What is the numerator for this measure? 8. Is this measure only valid only for comparing costs within the same well-defined population? 9. How does the use of the attribution guideline impact the calculation of the total cost index? Developer Response:

1. HealthPartners historically used a health plan -specific survey, but in the Minnesota community, they use Minnesota Community Measurement in the Minnesota community.
2. Benchmarking is done based on the plan average, so the variation for a health plan, for example, would be among the groups within the plan.
3. Comparisons between regions would be based on the ability to access an adequate data set, the type of attribution model that has been used employed by the measure user of the measure, and the business application of the measure (e.g., use by consumers or internal benchmarking).
4. The measure counts what the plan is paying, plus the member liability (i.e., member co-pay).
5. Medical and pharmacy PMPM costs are calculated separately and then added together. However, in the HealthPartners system there are no carve-outs for behavioral health. For systems that do have behavioral and pharmacy carve-outs, it is recommended that the user is consistent in how these data are cleaned and used in the measure.
6. For pharmacy costs, for example, the numerator would be the plan and plus the member co-pay, with the denominator being only those with the pharmacy benefit, thus accounting for the carve-out. Members impacted affected by the carve-out are not left out of the measure, but are examined separately with medical and behavioral together. They are accounted for at the aggregate level.
7. Total costs for patients in the group (100% of services), regardless of attribution rules.
8. It is possible to compare across these groups, but the measure would be used to show a cost differential. The user would then have the option of using a geographic adjuster to account for these differences in business applications.
9. Attribution does not impact the calculation of the index.

1.Importance to Measure and Report
1a.High Impact: H-15, M-2, L-1, I-0, N/A-0
Committee Discussion: The Committee agrees this criterion was adequately met.
1b. Resource use/cost problems: H-14, M-4, L-0, I-0, N/A-0
Committee Discussion: The Committee agrees this criterion was adequately met.
1c. Purpose clearly described: H-11, M-7, L-0, I-0, N/A-0
Committee Discussion: The Committee agrees this criterion was adequately met.
1d. Resource use service categories consistent and representative: H-11, M-7, L-0, I-0, N/A-0
Committee Discussion: The Committee agrees this criterion was adequately met.

1.Overall Importance: Y- 18, N-0

2.Scientific Acceptability of Measure Properties:
2a. Reliability:
2a1. Well- defined/precise specifications: H-5, M-8, L-4, I-1, N/A-0
Committee Discussion: There was concern that whether the total cost PMPM measurement for a health plan is useful, because it does not use standardized prices, it does not seem to be generalizable to different populations outside of the geographical region in which it is used. While geographic adjusters are available for helping to address regional differences, it should not be up to the user to figure this out along with the many other factors that contribute to the PMPM resource use/costs in a community. There was disagreement among the Committee on whether the lack of nationally comparability and potential limited use for this measure conflicts with the intent of endorsement. While some believe endorsed measures should be generalizable for various regions and markets, others believe it is useful and acceptable to have a measure endorsed for use within the context of a region for comparisons. There are some systems, health plans, and consumers that are interested in knowing actual costs. For example, there are many health systems are looking for this type of measure; particularly in California, for Medicare and commercial population ACO's, actual costs for total cost of care are of great interest. This measure provides real economic information that resource use measures that use standardized prices do not give information that will guide people's choices. If, for example, from an ACO's perspective, adjusting is undesirable, the actual total cost to the system is of interest for accountability purposes. The Committee and developers also acknowledged that all endorsed measures are not useful for every region and population.
2a2. Reliability testing: H-10, M-4, L-4, I-0, N/A-0
Committee Discussion: An analysis of the reliability testing was conducted by the NQF statistical consultant and shared with the Committee. His analysis was based on bootstrapping simulations restricted to each provider group; this was done three times in each year of data for each provider group. They used a variation simulation and compared its results to the observed variability to measure the signal-to-noise ratio. In addition, they compared how the ratios changed over time by provider, demonstrating insignificant differences. The reliability testing was deemed accepted and demonstrated a high level of measure score reliability.
Overall Reliability: H-8, M-6, L-4, I-0, N/A-0

2b. Validity:
2b1. Specifications consistent with resource use/cost problem: H-4, M-5, L-9, I-0, N/A-0
Committee Discussion: Committee members expressed a great deal of concern about the primary care attribution guideline submitted for this measure. Attribution instructions could be submitted as well thought- out guidelines, allowing for user flexibility to use the method outlined, or another method that suits the user's specific application while still enabling the use of the core measure specifications that have been validated. Developers also had the option of submitting attribution instructions as specifications, which require the user to apply the method specified in order to fully implement the measure fully. The attribution approach for this measure was submitted as guidelines. Within the context of these attribution guidelines, there were concerns with the inclusion of inpatient costs to the total cost, but the attribution model attributes based on outpatient resource use. For example, a doctor could be held responsible for a patient's inpatient stay before ever seeing the patient in an outpatient visit. There were concerns about how the use of this type of model might affect practice and potentially incentivize providers not to take on new patients who haven't have not seen a PCP. Another concern with the attribution guideline is accounting for care managed primarily by a specialist, since the guideline attributes to primary care providers (PCPs). Within the HealthPartners system 75% of its users use PCPs; this is not the case for many other areas in the country. Finally, within this attribution approach, non-users of the system are not attributed. This measure can be used in conjunction with measure 1598, which is specified in the exact same manner but uses standardized pricing. When used together the difference between the actual and standardized prices can be used to reflect differences in regional pricing.
Secondary to the concerns around the attribution guideline, is the level of analysis, which includes the physician group level. A physician group is defined by the developer as 2 or more physicians, with a recommended minimum of 600 patients in the sample.
The Committee voted on this criterion with the understanding that the attribution approach was submitted as a guideline.
2b2. Validity testing: H-7, M-5, L-5, I-0, N/A-0
Committee Discussion: The NQF Statistical Consultant conducted an analysis of the validity testing and shared it with the Committee. The validity testing sought to demonstrate face validity. Testing was conducted on provider groups, not for individual providers. As previously mentioned, the recommended minimum sample size is 600 patients. The Committee There expressed concern expressed about how would this measure operate for groups with only 2 -3 physicians.
2b3. Exclusions: H-3, M-6, L-9; I-0
Committee Discussion: Patients that who do not have a PCP are excluded from the denominator. The Committee expressed concern with this exclusion, as members who seek care from a specialist may be using resources within the system, but those resources are not counted in the total cost. This brings concerns that there may be potential for "gaming the system" using this measure—a system's total cost may appear lower if most of its care is provided by specialist. The issue of pharmacy carve-outs and how they are handled in this measure were also were discussed relevant to this criterion.
2b4. Risk-adjustment: H-7, M-7, L-2, I-2, N/A-0
Committee Discussion: This measure uses ACG's to risk adjust. It is a widely known and accepted methodology developed and maintained by a John's Hopkins group. The use of the ACG risk adjuster is open to the public for a fee based on the type of user. Fees associated with the using of the adjuster are discussed below in Feasibility criterion 4d. Adjustment in the underlying populations also has also been applied. The NQF Statistical Consultant conducted an analysis of the risk -adjustment model was conducted by the NQF statistical consultant and shared it with the Committee. The risk- adjustment model was included in a correlation analysis with the physician total cost index (TCI) and the observed actual costs, and which demonstrated that the risk -adjustment model adequately accounts for variation, lowering the correlation between the TCI and actual costs.
2b5. Identification of statistically significant/meaningful differences: H-7, M-5, L -2, I-4
Committee Discussion: Most Committee members agreed the measure adequately demonstrated this criterion. Others believed that given the concerns with the exclusions, focus on primary care encounters, validity testing at the group level only, and comparisons across regions, the ability to determine statistically significant differences is unclear.
2b6. Multiple data sources: N/A
Committee Discussion: N/A
Overall Validity: H-4, M-6, L-7, I-0, N/A-0
2c. Stratification for disparities: H-1, M-8, L-3, I-7, N/A-0
Committee Discussion: Due to the limitations in the administrative claims data to capture race and ethnicity, it is difficult to assess how they might be accounted for in the measure. However, if the data were available, the Committee agrees the measure is constructed such that it would be able to report stratified data. The HealthPartners system does collect race and language information and is working on eliminating disparities in its system; however, this measure has not been stratified to report on disparities at this time.

2.Overall Scientifically Acceptable: Yes [Y-9, N-10 (Committee Vote)]

3. Usability:

3a. Measure performance results are publicly reported: H-9, M-7, L-0, I-0, N/A-0 *Committee Discussion:* Measure is currently in use in the Minnesota region. 3b. Measure results are meaningful/useful for accountability and quality improvement: H-4, M-8, L-4, I-0, N/A-0 *Committee Discussion:* The Committee's discussion of the generalizability and comparability of the measure geographically and across varied patient populations also applies to the utility of this type of data for accountability and for the intended audiences. See discussion in 2a1, 2b1, and 2b2. 3c. Data and results can be decomposed for transparency and understanding: H-7, M-6, L-3, I-0, N/A-0 *Committee Discussion:* Behavioral health and pharmacy carve- outs are a concern. Comparisons should not be made between entities with carve- outs and those without. 3d. Harmonized or justification for differences: N/A Discussion: N/A
3.Overall Usability: H-6, M-7, L-2, I-0, N/A-0
4. Feasibility: 4a. Data elements routinely generated during care process: H-11, M-7, L-0, I-0, N/A-0 *Committee Discussion*: The Committee agreed this criterion has been adequately demonstrated as this measure uses administrative claims data, which are generated as a byproduct of care delivery. 4b. Data elements available electronically: H-11, M-6, L-1, I-0, N/A-0 *Committee Discussion*: The Committee agreed this criterion has been adequately demonstrated, as this measure uses administrative claims data, which are available electronically. Due to the issue of carve-outs, however, not all data are available electronically (i.e; pharmacy data). 4c. Susceptibility to inaccuracies/ unintended consequences identified: H-4, M-6, L-8, I-0, N/A-0 *Committee Discussion:* The committee suggested a title change to indicate this measure should only be used for measuring costs in the in primary care setting. Setting the threshold of a visit with the PCP should be more than 1 visit (HP responded nonusers can be brought into play at the health plan level). 4d. Data collection strategy can be implemented: H-0, M-13, L-3, I-0, N/A-0 *Committee Discussion*: Consideration of pricing table. Carve- outs an issue
4. Overall Feasibility: H-3, M-8, L-7, I-0, N/A-0

WITHDRAWN BY DEVELOPER

The measures listed below were withdrawn from the Cycle one review process by the developers for further refinement and testing.

- (1570) Acute Myocardial Infarction Episode-of-Care for 30 Days Following Onset (ABMS)
- (1571) Acute Myocardial Infarction Episode-of-Care for Post-Acute Period (days 31-365) (ABMS)
- (1572) Episode of Care for Management of Chronic Coronary Artery Disease (ABMS)
- (1573) Episode of Care for Management of coronary Artery Disease Post Re-Vascularization (ABMS)
- (1574) Episode of Care for Management of Chronic Congestive Heart Failure over a 12 month period (ABMS)

- (1575) Episode of Care for Management of Post-Hospitalization Chronic Congestive Heart Failure over a 4 Month Period (ABMS)
- (1576) Episode of Care for Patients with Diabetes over a One Year Period (ABMS)
- (1593) ETG Based Acute Myocardial Infarction (AMI) Resource Use Measure (Ingenix)
- (1596) ETG Based Stroke Resource Use Measure (Ingenix)

ADDITIONAL RECOMMENDATIONS

Recommendations and further guidance from the Committee on the applying of the endorsed measures and future resource use measurement efforts will be discussed in a subsequent report for Cycle two of this project.

NOTES

1. Catlin A, et al., National Health Spending in 2006: A Year of Change for Prescription Drugs, Health Affairs. 2008; 27(1):14–29.
2. Banks J, et al., Disease and disadvantage in the United States and in England, Journal of the American Medical Association. 2006; 295(17):2037–2045.
3. Hoyert DL, et al., Annual summary of vital statistics: 2004; Pediatrics 2006; 117(1):168–183.
4. Weiss JE, Mushinski M., International mortality rates and life expectancy: selected countries, Statistical Bulletin—Metropolitan Life Insurance Company. 1999; 80(1):13–21.
5. Catlin A, et al.
6. CMS Grouper Episode for Medicare. Federal Business Opportunities. Obtained from: https://www.fbo.gov/index?s=opportunity&mode=form&id=452b7af0f5752ba0c444853793e8024d&tab=core&_cview=1. Last Accessed August 2011.
7. Institute of Medicine (IOM), Crossing the Quality Chasm: A New Health Systems for the 21st Century, Washington, DC: National Academies Press; 2001.
8. AQA Principles of "Efficiency" Measures April 2006. Available from: http://www.aqaalliance.org/performancewg.htm. Last accessed August 2011.
9. Atlas, R. F., The role of PBMs in implementing the Medicare prescription drug benefit. Health Affairs. 2004; Jul-Dec; Suppl Web Exclusives: W4-504-15. Review.

10. Grazier, K. L. and L. L. Eselius., Mental health carve-outs: Effects and implications. Medical Care Research and Review. 1999; 56 (suppl 2): 37.

11. Hasings, C., Mosteller, F., Tukey, J.W., Winsor, C.P. Low moments for small samples: a comparative study of order statistics. Annals of Mathematical Statistics. 1947; 18, 413–426.

APPENDIX A—SPECIFICATIONS FOR COST AND RESOURCE USE MEASURES 2011 (Cycle 1)

The following tables present the detailed measure specifications for the recommended consensus standards. All information presented here has been derived directly from the measure developers without modification or alteration (except where measure developers agreed to such modifications) and is current as of August 15, 2011. All proposed voluntary consensus standards are open source, meaning they are fully accessible and disclosed.

	1557: Relative Resource Use for People with Diabetes (RDI)
Steward	National Committee for Quality Assurance (NCQA), | 1100 13th Street NW, STE 1000, | Washington, | District Of Columbia, 20005
Description	The risk-adjusted relative resource use by health plan members 18-75 years of age who were identified as having diabetes (type 1 and type 2) during the measurement year.
Resource Use Measure Type	Per capita (population- or patient-based)
Data Source	Administrative claims, Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data : Laboratory, Electronic Clinical Data : Pharmacy, Electronic Clinical Data : Registry, Paper Records NCQA collects HEDIS RRU data directly from Health Plan Organizations and Preferred Provider Organizations via a data submission portal - the Interactive Data Submission System (IDSS).). RRU measures use NCQA's standardized prices and NCQA collects data with only the standardized prices applied.
Level of Analysis	Specifications: Clinician : Group/Practice, Clinician : Individual, Clinician : Team, Facility, Health Plan, Integrated Delivery System, Population : Community, Population : County or City, Population : National, Population : Regional, Population : states
Costing Method	RRU measures use NCQA's standardized prices. The organization does not report prices based on its contracts and fee schedules, rather it applies a standard price to each service, multiplies it by the number of

	1557: Relative Resource Use for People with Diabetes (RDI)
	units of service and reports the resulting standard cost. Using this approach protects proprietary fee schedules and contracts while supporting equitable measure comparison across organizations and across regions without requiring adjustment for levels of service payment.
Tested Population	Commercial; Medicaid; Medicare
Resource Use Service Categories	Inpatient services: Inpatient facility services; Inpatient services: Evaluation and management; Inpatient services: Procedures and surgeries; Inpatient services: Imaging and diagnostic; Inpatient services: Lab services; Inpatient services: Admissions/discharges; Ambulatory services: Outpatient facility services; Ambulatory services: Emergency Department; Ambulatory services: Pharmacy; Ambulatory services: Evaluation and management; Ambulatory services: Procedures and surgeries; Ambulatory services: Imaging and diagnostic; Ambulatory services: Lab services
Attribution Approach	Specifications: Using administrative claims data submitted by all organizations, NCQA estimates the expected RRU amounts for each clinical condition for each organization. RRU index amounts are based on the ratio of observed to expected amounts. Results can be assessed at an overall basis, across all members and major clinical conditions, by service category or for a member cohort within a condition. Relative resource use is calculated at the plan-level and no attribution of resource use is made below this level. Attribution of resource use to a particular NCQA submission is based on the product line and reporting type of the plan that the member was enrolled in as of the end of the measure year.

796
797

	1558: Relative Resource Use for People with Cardiovascular Conditions
Steward	National Committee for Quality Assurance (NCQA), | 1100 13th Street NW, STE 1000, | Washington, | District Of Columbia, 20005
Description	The risk-adjusted relative resource use by health plan members with specific cardiovascular conditions during the measurement year.
Resource Use Measure Type	Per capita (population- or patient-based)
Data Source	Administrative claims, Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data : Laboratory, Electronic Clinical Data : Pharmacy, Paper Records NCQA collects HEDIS RRU data directly from Health Plan Organizations and Preferred Provider Organizations via a data submission portal - the Interactive Data Submission System (IDSS). RRU measures use NCQA's standardized prices and NCQA collects data with only the standardized prices applied.
Level of Analysis	Specifications: Clinician : Group/Practice, Health Plan, Integrated Delivery System, Population : National, Population : Regional
Costing Method	RRU measures use NCQA's standardized prices. The organization does not report prices based on its contracts and fee schedules, rather it applies a standard price to each service, multiplies it by the number of units of service and reports the resulting standard cost. Using this approach protects proprietary fee schedules and contracts while supporting equitable measure comparison across organizations and across regions without requiring adjustment for levels of service payment.
Tested Population	Commercial; Medicaid; Medicare
Resource Use Service Categories	Inpatient services: Inpatient facility services; Inpatient services: Evaluation and management; Inpatient services: Procedures and surgeries; Inpatient services: Imaging and diagnostic; Inpatient services: Lab services; Inpatient services: Admissions/discharges; Ambulatory services: Outpatient facility services;

	1558: Relative Resource Use for People with Cardiovascular Conditions
	Ambulatory services: Emergency Department; Ambulatory services: Pharmacy; Ambulatory services: Evaluation and management; Ambulatory services: Procedures and surgeries; Ambulatory services: Imaging and diagnostic; Ambulatory services: Lab services
Attribution Approach	Specifications: Using administrative claims data submitted by all organizations, NCQA estimates the expected RRU amounts for each clinical condition for each organization. RRU index amounts are based on the ratio of observed to expected amounts. Results can be assessed at an overall basis, across all members and major clinical conditions, by service category or for a member cohort within a condition. Relative resource use is calculated at the plan-level and no attribution of resource use is made below this level. Attribution of resource use to a particular NCQA submission is based on the product line and reporting type of the plan that the member was enrolled in as of the end of the measure year.

798
799

	1598: Total Resource Use Population-based PMPM Index
Steward	HealthPartners, | 8170 33rd Avenue South, PO Box 1309, | Bloomington, MN, 55425
Description	The Resource Use Index (RUI) is a risk adjusted measure of the frequency and intensity of services utilized to manage a provider group's patients. Resource use includes all resources associated with treating members including professional, facility inpatient and outpatient, pharmacy, lab, radiology, ancillary and behavioral health services.
Resource Use Measure Type	Per capita (population- or patient-based)
Data Source	Administrative claims, Other: Users administrative claims data base, Risk-adjustment Tool, Johns Hopkins ACG System Version 9.0, Standardized costing code table, Total Care Relative Resource Values (TCRRV) specification provided
Level of Analysis	Clinician : Group/Practice; Population : Community
Costing Method	Description: The Total Care Relative Resource Values (TCRRVs) are a grand linear scale of relative values designed to evaluate resource use across all types of medical services, procedures and places of service. The values are independent of price and can be used to evaluate providers, hospitals, physicians and health plans against their peers on their efficiency of resource use in treating like conditions. General Overview of Application: The TCRRVs are applied at the procedure level for each component of care with the exception of inpatient, which is applied at the full admission level. There is a TCRRV lookup table for each component of care where each claim's procedure is matched with the corresponding value. The TCRRV weights that are applied to the claim is tested for accuracy and a total TCRRV is calculated. The final step is to calibrate the total TCRRVs to the paid ratio between components of care using the paid adjustment factor. http://www.healthpartners.com/files/56500.pdf OR www.healthpartners.com/tcoc.
Tested Population	Commercial
Resource Use Service Categories	Inpatient services: Inpatient facility services; Inpatient services: Evaluation and management; Inpatient services: Procedures and surgeries; Inpatient services: Imaging and diagnostic; Inpatient services: Lab services; Inpatient services: Admissions/discharges; Inpatient services: Labor (hours, FTE, etc.); Ambulatory services: Outpatient facility services; Ambulatory services: Emergency Department; Ambulatory services: Pharmacy; Ambulatory services: Evaluation and management; Ambulatory services: Procedures and surgeries; Ambulatory services: Imaging and diagnostic; Ambulatory services: Lab services; Ambulatory services: Labor (hours, FTE, etc.); Durable Medical Equipment (DME)

	1598: Total Resource Use Population-based PMPM Index
Attribution Approach	Guidelines: To determine which members to include in the Total Resource Use measure, there are several options available depending upon your business purpose and unit of measure. If the unit of measure is an entire health plan or employer group, all members will be included in the Total Resource Use measure. If the unit of measure is a provider and members are required to select a primary care provider, we recommend using the member selected provider. When the member is not required to select a primary care provider, we recommend the use of an attribution algorithm to identify the member's primary care provider. The measure was tested using this methodology.

800
801

	1604: Total Cost of Care Population-based PMPM Index
Steward	HealthPartners, | 8170 33rd Avenue South, PO Box 1309, | Bloomington, MN, 55425
Description	Total Cost of Care reflects a mix of complicated factors such as patient illness burden, service utilization and negotiated prices.
Resource Use Measure Type	Per capita (population- or patient-based)
Data Source	Administrative claims, Other: Users administrative claims data base, Risk-adjustment Tool, Johns Hopkins ACG System Version 9.0,
Level of Analysis	Guideline: Clinician : Group/Practice, Population : Community
Costing Method	The Total Cost of Care considers 100% of health care services in the Total Cost Index and is calculated on a risk-adjusted paid per member per month basis as well benchmarked to a peer group. The paid amount (i.e., allowed) is inclusive of both plan and member liability.
Tested Population	Commercial
Resource Use Service Categories	Inpatient services: Inpatient facility services; Inpatient services: Evaluation and management; Inpatient services: Procedures and surgeries; Inpatient services: Imaging and diagnostic; Inpatient services: Lab services; Inpatient services: Admissions/discharges; Inpatient services: Labor (hours, FTE, etc.); Ambulatory services: Outpatient facility services; Ambulatory services: Emergency Department; Ambulatory services: Pharmacy; Ambulatory services: Evaluation and management; Ambulatory services: Procedures and surgeries; Ambulatory services: Imaging and diagnostic; Ambulatory services: Lab services; Ambulatory services: Labor (hours, FTE, etc.); Durable Medical Equipment (DME)
Attribution Approach	Guidelines: To determine which members to include in the Total Resource Use measure, there are several options available depending upon your business purpose and unit of measure. If the unit of measure is an entire health plan or employer group, all members will be included in the Total Resource Use measure. If the unit of measure is a provider and members are required to select a primary care provider, we recommend using the member selected provider. When the member is not required to select a primary care provider, we recommend the use of an attribution algorithm to identify the member's primary care provider. The measure was tested using this methodology.

802

APPENDIX B—COMMITTEE

National Voluntary Consensus Standards for Resource Use Committee

Tom Rosenthal, MD (Co-Chair)
UCLA School of Medicine, Los Angeles, CA

Bruce Steinwald, MBA (Co-Chair)
Independent Consultant, Washington, DC

Paul G. Barnett, PhD
VA Palo Alto Health Care System, Menlo Park, CA

Jack Bowhan
Wisconsin Collaborative for Healthcare Quality, Middleton, WI

Jeptha P. Curtis, MD
Yale University School of Medicine, New Haven, CT

Kurtis S. Elward, MD, MPH
Family Medicine of Albemarle, Charlottesville, VA

William E. Golden, MD
Arkansas Medicaid, Little Rock, AR

Lisa M. Grabert, MPH
American Hospital Association, Washington, DC

Ethan A. Halm, MD, MPH
University of Texas Southwestern Medical Center, Dallas, TX

Ann L. Hendrich, RN, MSN, PhD(c)
Ascension Health, St. Louis, MO

Thomas H. Lee, MD
Partners HealthCare System, Inc., Boston, MA

Jack Needleman, PhD
University of California, Los Angeles School of Public Health

Mary Kay O'Neill, MD, MBA
CIGNA HealthCare, Seattle, WA

David F. Penson, MD, MPH
Vanderbilt University Medical Center, Nashville, TN

Doris Peter, PhD
Consumer Reports, Yonkers, NY

Steve Phillips, MPA
Johnson & Johnson Health Care Systems Inc., Washington, DC

David Redfearn, PhD
WellPoint, Las Vegas, NV Woodland Hills, CA

Jeffrey B. Rich, MD
Mid-Atlantic Cardiothoracic Surgeons Ltd., Norfolk, VA

William L. Rich, III, MD
Northern Virginia Ophthalmology Associates, Falls Church, VA

Barbara A. Rudolph, PhD, MSSW
The Leapfrog Group, Fitchburg, WI

Joseph Stephansky, PhD
Michigan Health & Hospital Association, Lansing, MI

James N. Weinstein, DO, MS
The Dartmouth Institute for Health Policy and Clinical Practice & The Dartmouth-Hitchcock Clinic, Lebanon, NH

Dolores Yanagihara, MPH
Integrated Healthcare Association, Oakland, CA

NQF Staff

Helen Burstin, MD, MPH
Senior Vice President, of Performance Measures

Heidi Bossley, MBA, MSN
Vice President, of Performance Measures

Taroon Amin, MA, MPH
Senior Director

Ashlie Wilbon, RN, MPH
Senior Project Manager

Lauralei Dorian
Project Manager

Sarah Fanta
Project Analyst

APPENDIX C—CARDIOVASCULAR/DIABETES TECHNICAL ADVISORY PANEL

Jeptha Curtis, MD, FACC (Co-Chair)
Yale University School of Medicine, New Haven, CT

James Rosenzweig, MD (Co-Chair)
Boston Medical Center and Boston University School of Medicine, Boston, MA

Mary Ann Clark, MHA
Neocure Group, Washington, DC

Constance Hwang, MD, MPH
Resolution Health, Inc., Columbia, MD

Thomas Marwick, MBBS, PhD
Cleveland Clinic, Cleveland, OH

Michael O'Toole, MD
Midwest Heart Specialists, Ltd., Downers Grove, IL

David Palestrant, MD
Cedars-Sinai Medical Center, Los Angeles, CA

Brenda Parker, PharmD
GlaxoSmithKline, Marietta, GA

Katherine Reeder, PhD, RN
University of Kansas School of Nursing, Kansas City, KS

William Weintraub, MD
Christiana Care Health System, Newark, DE

APPENDIX D—RESOURCE USE MEASUREMENT TERMS

The following resource use measurement terms have been defined based on their use in the context of this project and are important to understanding the concepts in this report.

Attribution- identifying and assigning of a responsible provider or entity (e.g., health plan) for the care delivered for an episode or population.

Benchmarking - the process of comparing the performance of accountable entities with that of their peers or with external best practice results. In developing comparative estimates, results should be risk adjusted for patient-level attributes to support the valid comparisons of these accountable entitles.

Carve-outs - the outsourcing of services, such as behavioral health or pharmacy claims, to specialty health plans or claims processing entities or organizations.

Clinical hierarchy - an arrangement of clinical conditions that are ranked according to severity, as "high," "below," or "at the same level." For example, if a patient has COPD and develops bronchitis, COPD would be assigned a greater weight than bronchitis.

Exclusion criteria - criteria applied before a measure is tested in order to remove any individuals with conditions that may skew the final measure score.

Peer groups - the ways in which resource use measures ensure providers and health plans are compared to similar providers and health plans.

Per capita measure – counts all services provided to a person within a specific population, regardless of condition or encounters with system.

Per episode measure - counts resources based on bundles of services that are part of a distinctive event provided by one or multiple entities (e.g., health services provided associated with an event or series of events for acute myocardial infarction)

Resource use service categories - categories of resource units or services provided care for a patient or population. Resource units are generally are identified through claims data and grouped into categories with similar types of claims (e.g., x-rays grouped into imaging category). Categories are generally are and measured in terms of dollars, but also can also include resources not captured on a claim (e.g., nursing hours).

Risk-adjustment - a corrective approach designed to reduce any negative or positive consequences associated with caring for patients of higher or lower health risk or propensity to require health services.

Severity levels - pre-determined levels of acuity used to rank and assign patients based on an assessment of the aggregate of their conditions/diagnosis codes.

Standardized pricing - pre-established uniform price for a service, typically based on historical price, replacement cost, or an analysis of completion in the market; removes variation in resource costs due to differences in negotiated prices or geographic differences based on labor or other input costs.

Stratification - division of a population or resource services into distinct, independent strata, or groups of similar data, enabling analysis of the specific subgroups. This type of adjustment can be used to show where disparities exist or where there is a need to expose differences in results.