# 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 2): A Consensus Report*

DA: October 20, 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.

Eleven measures were evaluated for suitability as voluntary consensus standards for accountability and performance improvement during review cycle two. 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. To date, the Steering Committee has recommended a total of eight cost and resource use measures for endorsement; four of which are a result of the second review cycle. For one measure (1595: ETG based diabetes cost of care measure), the Steering Committee vote was split and there was no consensus. The Committee seeks public and member comment on this measure to be considered in subsequent Committee discussion to determine the final recommendation for the measure.

The draft document, *National Voluntary Consensus Standards for Cost and Resource Use (Cycle 2): 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.

This report builds upon the commenting draft report posted for cycle 1. Elements of the report that have been enhanced and added to the draft are reflected in blue text. Following the measure summaries, a new section has been added to the report entitled "Additional Considerations" In addition to commenting on the measures, we request that comments also be submitted on the "Applying Resource Use Measure Evaluation Criteria" and "Additional Considerations" sections of the report. Please submit your comments on these sections 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.

# NATIONAL QUALITY FORUM

**NQF Member comments must be submitted no later than 6:00 pm ET, November 21.**
**Public comments must be submitted no later than 6:00 pm ET, November 14.**

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 2): A CONSENSUS REPORT

**DRAFT REPORT FOR COMMENTING**

**OCTOBER 20, 2011**

# NATIONAL QUALITY FORUM

## NATIONAL VOLUNTARY CONSENSUS STANDARDS FOR COST AND RESOURCE USE (CYCLE 2): A CONSENSUS REPORT

## TABLE OF CONTENTS

## NATIONAL VOLUNTARY CONSENSUS STANDARDS FOR COST AND RESOURCE USE (CYCLE 2): 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 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 measures of 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 11 measures reviewed during the second review cycle.

In the second cycle of the project, four additional measures have been recommended for endorsement as voluntary consensus standards suitable for accountability and performance improvement:

- (1560) Relative resource use for people with asthma (NCQA)
- (1561) Relative resource use for people with COPD (NCQA)
- (1609) ETG based hip/knee replacement cost of care measure (Ingenix)
- (1611) ETG based pneumonia cost of care (Ingenix)

## NATIONAL VOLUNTARY CONSENSUS STANDARDS FOR COST AND RESOURCE USE (CYCLE 2): A CONSENSUS REPORT

## BACKGROUND

The United States' healthcare 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 decreased mortality, increased patient satisfaction, or led to improvements in access or higher quality of 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' healthcare spending continues to increase at a rate of seven percent per year and is largely focused on treating acute and chronic illness rather than preventive 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. Linking resource use (or cost) and quality measures 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 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 second phase of a two-phase effort will endorse resource use measures through the Consensus Development process (CDP). These measures will serve as building blocks for efficiency of care measures and signal to the measure development industry 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 to evaluate 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 14 focus areas were assigned:

**Cycle 1**

**Cardiovascular**

- Congestive heart failure (CHF)
- Coronary artery disease (CAD)
- Acute myocardial infarction (AMI)

**Stroke**

**Diabetes**

**Non-condition specific** (e.g. per capita-population)

**Cycle 2**

**Pulmonary**

- Chronic obstructive pulmonary disease (COPD)
- Asthma
- Pneumonia

**Cancer**

- Breast cancer
- Colorectal cancer

**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 for 11 measures reviewed during Cycles 1 and 2.

## 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 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 healthcare spending, and when paired with quality measures, can help evaluate the efficiency of the healthcare 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 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 healthcare 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.

Given the diverse perspectives on cost and resource use measurement in healthcare, it is important to articulate, in the context of this project and the measures submitted, the terminology, purpose, and perspectives these measures represented. Recognizing this is NQF's first project in the resource use measurement arena, there is a clear gap in the NQF portfolio for these types of measures. NQF also recognizes that while the measure submission process is open to any entity wishing to submit measures for evaluation, the measures submitted and evaluated in this process are not representative of all approaches to measuring healthcare costs and resources that exist in the market today. This report is a reflection of the evaluation process of the measurement approaches submitted to this project for review.

Each of the measurement approaches submitted for review calculate the use of various resources using administrative claims data, categorize them by type of resource [e.g., pharmacy, durable medical equipment, evaluation and management (E&M) visits] and apply a costing methodology (either actual prices paid or standardized prices). When developers further apply a dollar value to utilization counts, the dollar value serves as a weight for each resource. Due to the limitations in the data types available for measuring resource use in healthcare, administrative claims data are the primary source of this information for the measures submitted to this project. Further

discussion of costing approaches and the use of administrative claims data are addressed later in the report.

Also important to understand in the context of this report is the way in which the terms "cost," "resource use," and "prices" are used. The term "cost" can represent very different constructs to various stakeholders. In the context of this report, cost (or cost of care measures) reflects the actual prices *paid* by health plans for health plan member for utilization; resource use or "resource use measures" further apply standardized prices to utilization counts. Prices charged by providers in healthcare, by many accounts, is not a good measure of utilization as prices charged can be a reflection of the negotiating position of health plans vis-à-vis providers in a given market. Prices paid is generally a reflection of the cost the health plan incurs to cover the claims submitted for its members; some measures also report a member (consumer) cost based on member co-pays. For a provider, (e.g., a physician or nurse practitioner) a cost of care measure would reflect the payment the provider received from the health plan for care provided. For a purchaser, a resource use measure can be used to assess the utilization of healthcare services across health plans, while a cost of care measure can be used to assess how well a health plan is managing charges and utilization of providers within the health plan's network. Given the other types of costs attributed to healthcare, it is important to note that these measures do not capture or represent production costs (fixed or any other costs to the provider to deliver care), administrative costs, government funding to support healthcare delivery, or societal costs (e.g., lost wages, sick days).

## 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. Within the Cycle 2 condition areas, 19 measures were submitted and evaluated for suitability as voluntary consensus standards for accountability; 12 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 efforts to refine the evaluation criteria for resource use measures and evaluate resource use measures for endorsement:

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 result should be clear and understandable 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 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 its ratings and recommendations for multiple measures, recognizing the need to balance the quantity and specificity of information required to evaluate adequately 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 measure's focus (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, evaluating the scientific acceptability of resource use measures includes reviewing the measure's specifications, reliability and validity testing, and approach to addressing disparities. The completeness, repeatability of the specifications, and the adequacy of

the reliability testing methodology and results are evaluated within the reliability criterion. Applying the validity criteria, the Committee was asked to determine whether the specifications reflected the intent of the measure and address 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 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 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 comorbid 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 often are limited by the ability of administrative claims data to assess patient health status and severity accurately. For example, measures submitted were unable to differentiate between community-acquired and healthcare-acquired pneumonia. Measures submitted also were unable 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.

The Committee evaluated the validity of the measures by examining the interaction of the measure components including the specified level of analysis and the risk adjustment approach. There is a need for nationally endorsed measures at the individual clinician level of measurement and the experts encourage development of measures at this level. However, the Committee expected developers to demonstrate statistical differences at sample sizes that would be observed in the level of analysis specified. Further, attribution of the measure to the individual or group practice level was discussed at length, focusing on the appropriateness and generalizability. While sample size and attribution could be submitted as guidelines, the Committee agreed these testing results contribute to the measure's scientific acceptability.

Measures submitted as a part of an episode grouper were challenging to evaluate because the assignment of claims into the episode, comorbidities and interactions, clinical hierarchies, and the handling of concurrent of clinical events included lengthy algorithm decision trees that were at times unclear and only partially transparent to the reviewers. 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 were not clearly explained in the evaluation of single resource use measures. 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. The Committee expected developers to provide a clear and transparent explanation of this tie-breaker logic, how claims would be assigned to episodes, and how various open episodes interact with each other. While resource use measures are complex, developers have a responsibility to provide an explanation of the construction logic within the grouper; however the explanations submitted were often insufficient.

*Adjustments for comparability*

A measure's result can be influenced by confounding external factors that can affect the measure score. Measure developers submitted steps for adjusting 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 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 drives categorization into different risk categories and the goodness of fit of the risk model. Of the various methodologies reviewed, none was 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, as the Committee deemed this report to be an inadequate analysis of the risk-adjustment models for the purposes of this project. For example, the Committee asserted that risk-adjustment models be tested and may need to be recalibrated based on the measure's target population. Guidance presented in the SOA report was insufficient in assisting the Committee's assessment of risk-adjustment model performance across various datasets, across various homogenous populations (including Medicaid or Medicare), or the credibility of risk-adjustment models across various population sizes. The Committee agreed that submissions lacking the necessary information to evaluate the risk model fully should not be considered in future efforts to evaluate resource use measures. Descriptions of the risk models should include model calibration statistics (i.e., the R-squared value), a discussion of how variables were selected (i.e., based on statistical significance or clinical indicators), and sensitivity analyses.

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 prices 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. 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. The Committee agreed that both approaches could be appropriate for different applications; however a measure used as a national consensus standard must select a single costing approach. Including both costing approaches within the same measure could reduce comparability and limit the user's ability to identify the source of variation. For this reason, developers that submitted a single measure with an option for the user to determine which costing method to apply were asked either to split the submission into two separate measures or select one of the approaches to apply to a single measure submission. At the Committee's request, measures that were unknowingly evaluated and voted on with optional costing approaches were re-voted during the Cycle 2 Committee meeting based on developer selection of a single costing approach to be applied (actual prices paid) to all of their measures.

Subsequent Committee discussions on applying an actual price approach for national comparisons at an individual provider level identified additional concerns. Specifically, the Committee noted the potential for misinterpreting physician resource use in national reporting. This pricing approach includes environmental factors (i.e., local facility and labor costs) that may be outside of an individual provider's control. The Committee agreed that when actual prices

paid are reported, utilization counts should be reported as well. The concern over the use of actual prices also was considered in the measure's usability. However, there was agreement that actual prices paid by health plans to providers is important to measure and report; for example, regional comparisons at the individual provider level where environmental factors may not be as prominent, or nationally at higher levels of measurement (i.e. health plan level). Measures based on actual prices paid are encouraged for endorsement, noting that the validity will be examined through the interaction of the measure's specified level of analysis, risk adjustment model, and attribution approach.

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. While the Committee encouraged measurement at the individual and group practice level, measures submitted to this project had difficulty demonstrating reliability and validity at this level. Across all levels of measurement, the Committee engaged in a detailed evaluation of the risk adjustment approach and minimum sample size to ensure that the measures produced a valid and 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, allowing 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. 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 are specified at the individual provider level, confidence intervals need to be presented, especially when displaying information with small sample sizes. Using 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 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 95$^{th}$ percentile are set to the value at the 95$^{th}$ 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 affects 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 critically important 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, or that the measure score is precise. Validity testing must demonstrate that the measure data elements are correct 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 driven by the fact that the submitted resource use measures 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. While administrative claims data reduces measure error due to manual chart abstraction and transcription, developers cannot rely on the administrative claims to capture patient clinical characteristics accurately without proper data element validity testing. Claims data provide only limited clinical information, lack detail in determining patient health severity, and are subject to variation in coding processes by the accountable entities. The Committee agreed that these concerns span measures of quality and resource use and are not limited to the measures currently under evaluation.

*Measure score reliability*

Measure developers also performed varying levels of reliability assessments at the measure score level. The Committee was interested in assessing the measure's precision or ability to detect signal rather than noise. Measures demonstrated lower levels of measure score reliability assessments including parallel development of episode grouper software and SAS using the exact same specifications. While these tests demonstrated match rates of more than 99.9 percent, they do not facilitate assessments of the measure score's precision. Further, developers whose measures have been in use attempted to demonstrate the reliability of the observed/expected results (O/E) over time; however, doing so does not provide an assessment of precision of the measure score. The Committee suggested other robust methodologies that could be used to demonstrate a high level of reliability, including 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 because there were no comparisons to other independent claims databases or other authoritative data sources (e.g., the

patient's medical record). In addition, 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; ensuring 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 measures' 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 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 on 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 implementing resource use measures, which often use very complex, sophisticated methodologies to adjust risk and determine episode logic, for example. The TAPs and the Committee discussed this issue of complexity for the implementer (and for the users of the results) 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 marketplace 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.

## Harmonization and Best-in-Class

In Phase One of this resource use measurement project, the Committee agreed that because this is NQF's first effort focused on evaluating resource use measures, identifying "best-in-class" and requiring harmonization among resource use measures would be premature. In the context of resource use measures, similar measures may share the same measure type (e.g., per episode, per capita), or measure the same costs/resources (e.g., actual prices paid vs. standard prices, resource service categories), or address the same population (e.g., people with diabetes). Competing measures would share all of the characteristics previously listed. Among the eight measures recommended for endorsement, there were no competing measures. Recommended measures that were the same measure type were submitted from the same developer and were already harmonized. With the exception of the two non-condition-specific total cost of care measures (submitted by the same developer and recommended in Cycle 1), which employ different costing methodologies, all recommended measures addressed different populations. Future resource use measure endorsement efforts should explore the potential ways in which harmonization among similar measures might be achieved. Specifically, identifying which measure constructs (e.g.,

condition-specific episode trigger and end mechanisms, age ranges), if any, could be harmonized for standard measurement is needed in this measurement area. Also, exploring the implications of harmonization for the resource use measure development community in which proprietary measure components are common would be useful as the portfolio of endorsed resource use measures expands.

## RECOMMENDATIONS FOR ENDORSEMENT

This report presents the results from the evaluation of 11 measures considered during review cycles one and two under NQF's CDP.

### *Evaluation of Measure Costing Approaches*

Early in the evaluation process, the Committee agreed that it was important to distinguish measure results obtained using standardized prices and actual prices paid; dividing the costing approaches into separate measures was determined to be the best approach to ensure this distinction was made for standardized implementation and prevent inaccurate comparisons. As such, developers that submitted a single measure with an option for the user to determine which costing method to apply, were asked either to split the submission into two separate measures, or select one of the approaches to apply to a single measure submission. This was requested of both HealthPartners (in cycle one) and of Ingenix (in cycles one and two). HealthPartners subsequently resubmitted two separate measures, one applying each costing approach; Ingenix resubmitted all of their measures applying only actual prices paid.

During the initial evaluation and voting for recommendation of the Ingenix measures, there was not a shared understanding among the Committee that the measures had been submitted with flexibility in the costing approach. Ingenix chose to resubmit their measures using actual prices paid. Once the measures were resubmitted to the Committee applying the single costing approach, the Committee was given the opportunity to determine if the selection in the costing approach warranted a re-vote. The Committee requested a revote since there was not a shared understanding on the original costing approach by Ingenix, thus all Ingenix measures were re-

voted on during the Cycle 2 Committee meeting. The re-vote was for overall recommendation for endorsement only. This is reflected as such in the measure evaluation summaries below.

### *Evaluation of Measurement Approaches*

The NQF measure evaluation process calls for each submitted measure to be evaluated individually, based on its own merit. This was also the approach used in this project. Additionally in this project, given the nature of the various of resource use measure developers, measures developed by a single developer shared many common underlying measure constructs and processes. By understanding the common constructs shared among a group of measures from a developer (i.e. general methods), it lays the foundation for understanding the nuances specific to each individual measure. During the measure evaluation process, the TAPs and Committees often identified some recurring themes within the criteria discussions that applied across measures from an individual developer, regardless of condition focus of the individual measure. Some of these recurring themes have been captured in several of the measure evaluation summaries and some have been identified below.

*Ingenix Feasibility*

Each of the individual Ingenix measures [(Episode Treatment Groups (ETGs)] exist as part of a larger grouper system, and requires the use of the entire grouper to produce results for the individual ETGs. Because, each of the condition-specific ETGs submitted to this project require the use of the Ingenix grouper product to implement the measures, the Committee's discussion of the feasibility criterion for these measures was done for all of these measures at one time. As a part of feasibility discussion, the Committee was provided with a pricing table for each of the products required for implementation of these condition-specific ETGs.

Because these measures primarily use administrative claims data, all of the data required to implement these measures is generated as a byproduct of care and is available electronically. There was concern around the measure's susceptibility to inaccuracies as Ingenix does not have a formal audit system to ensure that all of data is included and correct. In terms of barriers to use,

the purchase and implementation of this product could be cost prohibitive for some entities. Annually, for physicians the cost to implement this project could range from of the small package $70,000 (for a group of less than 800 physicians) to $110,000 (for over 2,000 physicians in the group). For health plans, the annual cost could range from $90,115 (for less than 400,000 covered lives) to is $135,000 (for over a million covered lives). The Steering Committee concluded that this cost is comparable to the cost of other proprietary fees associated with other risk adjustment models of its caliber (e.g., ACGs used by HealthPartners). These prices include costs associated with the licensure of the proprietary software and the cost of all of their measures, over 558 ETGs, but not implementation. The Steering Committee acknowledged that while the methodology is very complex, the system may be used without Ingenix's technical support, if the user spends time thoroughly reviewing the documentation.

## 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:

Pulmonary



Bone/Joint

## Evaluation Summary—Candidate Consensus Standards Recommended for Endorsement

| 1560: Relative Resource Use for People with Asthma (NCQA) |
|---|
| Description: This measure addresses the resource use of members identified as having asthma. Both encounter and pharmacy data are used to identify members for inclusion in the eligible population, and the results are adjusted to account for age, gender, and HCC-RRU risk classifications that predict cost variability (Refer to Attachment S8_Clinical Logic for additional information).<br>Resource Use Type: Per capita (population- or patient-based)<br>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<br>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<br>Care Setting: Ambulatory Care : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Hospital/Acute Care Facility, Imaging Facility, Laboratory, Pharmacy, Post Acute/Long Term Care Facility: Nursing Home/Skilled Nursing Facility<br>Level of Analysis: Clinician: Group/Practice, Health Plan, Integrated Delivery System, Population : National, Population: Regional<br>Measure Developer: National Committee for Quality Assurance (NCQA), 1100 13th Street NW, STE 1000,Washington, District Of Columbia, 20005 |
| Committee Recommendation for Endorsement: Y-13; N-0; Abstain-1 |
| Conditions/Questions for Developer:<br>1. Could this measure be improved by including other diagnostic criteria to ensure all appropriate asthma patients are captured?<br>2. How have you come up with the age strata in your risk-adjustment?<br>3. Can secondary diagnosis be taken into account within the measurement year?<br>4. Is cost during the measurement year part of the risk-adjustment strategy?<br>5. Are your measure results published publically?<br>Developer Response:<br>1. Using asthma as a principal diagnosis will make it difficult to identify most patients, especially those who are acute and come into the ER and are diagnosed with bronchitis first, and then asthma.<br>2. The age strata for risk-adjustment are designed around known utilization patterns and clinical treatment patterns.<br>3. *All* costs for anyone with asthma are counted.<br>4. The HCC uses any services during the year to appropriately categorize patients into those 13 risk cohorts by severity of comorbidity. They also look at ICD-9 and procedural codes to categorize them and then go back and look at the number of times those services were offered to that population. Therefore, if a patient has multiple co-morbidities, that factors into the risk-adjustment, and will put a patient into a more severe risk-adjustment category.<br>5. Results are published through NCQA's Quality Compass module which contains the individual plan results by detailed service category along with a quality score. |
| 1. Importance to Measure and Report<br>1a.High Impact: H-9; M-0; L-0; I-0<br>*TAP Discussion:* The TAP agrees that asthma is an important area of healthcare to measure due to its high cost and the potential for improvements in care.<br>1b. Resource use/cost problems: H-7; M-2; L-0; I-0 |

*TAP Discussion:* The TAP agrees that asthma represents a resource use problem and noted that there is a well-documented opportunity for improvement.
1c. Purpose clearly described: H-9; M-0; L-0; I-0
*Discussion:* The TAP believes the purpose and objective are clear; this subcriterion has been met.
1d. Resource use service categories consistent and representative: H-9; M-0; L-0; I-0
*TAP Discussion:* The TAP believes this subcriterion has been met; there were not issues raised.

Overall Importance: Y-16, N-0
Committee Discussion: The Steering Committee agrees this criterion has been met.

2. Scientific Acceptability of Measure Properties:
2a. Overall Reliability: H-8; M-1; L-0; I-0
2a1.Measure well defined and precisely specified: H-9; M-0; L-0; I-0
*TAP Discussion*: The TAP believes this subcriterion has been met.
2a2. The results are repeatable: H-8; M-1; L-0; I-0
*TAP Discussion:* There was general agreement from the TAP that following a methodology of including *all* costs avoids having to consider what costs should or should not be associated with asthma. The developer reaffirmed that the measures are valid for any health plan; they are population-based measures and have been tested and can be used in physician groups with a sufficient number of patients. A population of at least 400 members is needed for the methodology to be valid, so it consequently tends to be larger physician groups that can use the measures.
2b. Overall Validity: H-5; M-4; L-0; I-0
2b1. Evidence is consistent with intent: H-6; M-3; L-0; I-0
*TAP Discussion*: The TAP agrees there is good overall evidence of face validity, but also a general desire to see more specific discussion around the face validity of the use of HCC's in this population.
2b2.Score/Analysis: H-6; M-3; L-0; I-0
*TAP Discussion:* The face validity of HCC's was found to be clear, but the logic behind the age stratification was unclear. 2b3. Exclusions: H-6; M-3; L-0; I-0
*TAP Discussion:* The TAP had an in-depth discussion regarding measure exclusions. The measure developer explained that cardiovascular conditions are not specifically excluded, but are used in the risk adjustment model. Patients with COPD are excluded. Exclusions affect the denominator population over either year within the two-year criteria, which is similar to the HEDIS asthma measure. There was agreement that the exclusion of COPD (which resulted in 38% of the initial population being eliminated) seems appropriate, particularly in light of the age range increasing to 64. The TAP did express concern that excluding acute respiratory failure could exclude poorly managed asthma patients. However, NCQA noted that acute respiratory failure only accounted for 3% of the population, so it doesn't meet their 5% threshold of concern.
2b4. Risk Adjustment: H-7; M-2; L-0 ; I-0
*TAP Discussion:* The TAP believes the risk-adjustment strategy seems appropriate. Several strategies are tested by NCQA, and the same methodology is used for all of their measures. The developer stratifies the population by age and gender and uses HCC's to risk adjust the population.
2b5. Identification of statistically significant/meaningful differences: H-8; M-1; L-0; I-0
*TAP Discussion:* There was general agreement that the distribution of the scores' detail score was appropriate. There was concern regarding whether the measure score could differentiate statistically significant and clinically significant variation.
2b6. Multiple data sources: N/A
2c. Stratification for disparities: H-5; M-3; L-0; I-1
*TAP Discussion:* The TAP believes stratification is needed although the data isn't available at this time.

Overall Scientifically Acceptable: Yes [Y-12; N-2 (Committee Vote)]
Overall Reliability: H-12; M-3; L-0; I-0
Overall Validity: H-4; M-9; L-1; I-0
Committee Discussion: The Committee agreed with the TAP's analysis of reliability and raised no additional concerns. There was further discussion around missing pharmacy data, and confirmation that plans submit separate components (total medical, quality, and pharmacy, for example) to NCQA and are allowed to have a certain number of missing components. NCQA then holds the plans accountable for ensuring that they have the complete data required to report the measure, and any plans that are missing a major component of the measure specification would not end up in the NCQA reporting product. The Committee asked the developers to defend the measure's use of indirect standardization in creating standardized prices.

Usability:

3a. Measure performance results are publicly reported: H-8; M-1; L-0; I-0
*TAP Discussion:* The TAP was satisfied that NCQA publically reports measure results and provides support to enable understanding of those results. Purchasers are using this information, along with NCQA quality measures, to improve value for their employees. Asthma is a bit more difficult because there is only one NCQA quality measure to associate with this cost measure, however there are more quality measures in the pipeline.
3b. Measure results are meaningful/useful for public reporting and quality improvement: H-6; M-3; L-0; I-0
*TAP Discussion:* The measure is straightforward and easy to interpret. NCQA uses standardized pricing tables, which are reviewed annually. Health plans are the main users for this data. However, purchasers and the large employers will also drive a need for this information. The TAP wondered how smaller businesses would implement this measure, and NCQA explained that they provide help through their annual conferences, webinar services and a dedicated webpage.
3c. Data and results can be decomposed for transparency and understanding: H-8; M-1; L-0; I-0
*TAP Discussion:* The TAP believes the methodology was transparent and appropriate.
3d. Harmonized or justification for differences: N/A

Overall Usability: H-9; M-5; L-0; I-0
Committee Discussion: The Steering Committee was concerned about the ability of small groups to implement this measure.

4. Feasibility:
4a. Data elements routinely generated during care process: H-9; M-0; L-0; I-0
*TAP Discussion*: The TAP agrees this subcriterion has been met; the data is a byproduct of care.
4b. Data elements available electronically: H-9; M-0; L-0; I-0
*TAP Discussion*: The TAP agrees this subcriterion has been met; the data is available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-7; M-2; L-0; I-0
*TAP Discussion*: There was agreement that NCQA did a sufficient job recognizing where the challenges with data inaccuracies are and have adequately addressed these challenges.
4d. Data collection strategy can be implemented: H-8; M-1; L-0; I-0
*TAP Discussion*: All the data submitted to NCQA must go through a certified auditor before it's reported to NCQA. As part of their annual analysis, NCQA reviews outliers, but currently the outliers are less than half a percent for this measure.

Overall Feasibility: H-10; M-4; L-0; I-0
Committee Discussion: No additional concerns were raised by the Steering Committee regarding feasibility.

867

1561: Relative Resource Use for People with COPD (NCQA)

Description: This measure addresses the resource use of members identified with COPD.
Clinical diagnosis of COPD during the measurement year is used to identify members for inclusion in the eligible population and the results are adjusted to account for age, gender, and HCC-RRU risk classifications that predict cost variability (Refer to Attachment S8_Clinical Logic for additional information).
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 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
Care Setting: Ambulatory Care : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Hospital/Acute Care Facility, Imaging Facility, Laboratory, Pharmacy, Post Acute/Long Term Care Facility: Nursing Home/Skilled Nursing Facility
Level of Analysis: Clinician : Group/Practice, Health Plan, Integrated Delivery System, Population: Community, Population: National, Population : Regional
Measure Developer: National Committee for Quality Assurance (NCQA), 1100 13th street NW, STE 1000,Washington, District Of Columbia, 20005

Committee Recommendation for Endorsement: Y-13; N-0; Abstain-1

Conditions/Questions for Developer:

1. If the goal is to eventually link these measures with quality measures and stratification is different, how will that be plausible?
2. What is the upper age limit to be included in this measure?
3. How do you ensure similar populations are compared?

Developer Response:

1. The resource use strata are different than they are for clinical quality strata, which are not risk-adjusted. As the quality measures further increase and perhaps in the future become risk-adjusted, there will be more room for comparability.
2. There is no upper age limit to this measure.
3. By risk adjusting to the specified level using the HCC's and the 13 different cohorts, NCQA end up comparing relatively similar plan populations. The quality index for this measure is use of diagnostic spirometer and exacerbations measures. There is no attribution of specific procedures to COPD yet.

1. Importance to Measure and Report

1a.High Impact: H-9; M-0; L-0; I-0

*TAP Discussion:* The TAP was in agreement that this is an important area of measurement.

1b. Resource use/cost problems: H-9; M-0; L-0; I-0

*TAP Discussion:* The TAP believes while there is variation in resource use was identified in other parts of the submission, the information submitted in the form for this item only discussed the variations in clinical care provided.

1c. Purpose clearly described: H-8; M-1; L-0; I-0

*TAP Discussion:* The TAP was concerned that the measure submission applied only to newly diagnosed patients. The developer clarified that it is supposed to apply to anyone with a diagnosis with COPD. Otherwise, the purpose of the measure is to evaluate the total cost of care for COPD patients within a 1 year timeframe was clear.

1d. Resource use service categories consistent and representative: H-9; M-0; L-0; I-0

*TAP Discussion:* The TAP believes this subcriterion has been met.

Overall Importance: Y-14, N-0

Committee Discussion: The Steering Committee agreed the measure focused on an important area of healthcare.

2. Scientific Acceptability of Measure Properties:

2a. Overall Reliability: H-7; M-2; L-0; I-0

2a1.Measure well defined and precisely specified: H-9; M-0; L-0; I-0

*TAP Discussion*: The TAP believes the specifications provided are clear and precise. The developer provided clarification on age stratification for resource use categories indicating that they are based on utilization patterns in the data-set, not clinical factors.

2a2. The results are repeatable: H-8; M-1; L-0; I-0

*TAP Discussion:* A similar methodology was used for this measure as for NCQA measure #1560, the primary difference being in the selection of the population. The TAP was concerned about the multiple populations being studied including commercial, Medicare, and Medicaid, due to the age range (unlike Measure 1560, where the age range cut off at 64). There was also concern that NCQA did not distinguish the fee-for-service versus the beneficiaries in Medicare Advantage plans.

2b. Overall Validity: H-4; M-5; L-0; I-0

2b1. Evidence is consistent with intent: H-8; M-1; L-0; I-0

*TAP Discussion*: The TAP believes the measure is clearly defined; however, one of the challenges will be the fact that COPD has multiple co-morbidities, particularly when compared to asthma. It will therefore be difficult to know if you are measuring exactly COPD. Specifications should be explored on how to develop disease severity; however, this is difficult to do with administrative datasets.

2b2.Score/Analysis: H-6; M-3; L-0; I-0

*TAP Discussion:* The TAP believes that overall the validity testing was appropriate. Outliers are identified by tagging O/E ratios below .3 or above 3.

2b3. Exclusions: H-4; M-5; L-0; I-0

*TAP Discussion:* The TAP agrees the exclusions are well stated and are similar to the asthma measure.

2b4. Risk Adjustment: H-6; M-3; L-0; I-0

*TAP Discussion:* Cardiovascular disease maybe a major driver of the severity of COPD.. The risk adjustment approach appears reasonable for the data available. The intent is to compare across populations.

2b5. Identification of statistically significant/meaningful differences: H-5; M-4; L-0; I-0

*TAP Discussion:* The TAP believes NCQA did a sufficient job presenting their data in a transparent manner.

2b6. Multiple data sources:

*TAP Discussion:* N/A (using all administrative data)

2c. Stratification for disparities: H-5; M-4; L-0; I-0

| |
|---|
| *TAP Discussion:* Examining differences in racial disparities for this data set is not yet possible, but there is stratification by gender. Race is not a required field for most provider systems and is usually unavailable except in the Medicare population. |
| Overall Scientifically Acceptable: Yes [Y-13; N-1 (Committee Vote)]<br>Overall Reliability: H-11; M-3; L-0; I-0<br>Overall Validity: H-4; M-10; L-0; I-0<br>Committee Discussion: The Steering Committee was satisfied by the appropriateness of the risk-adjustment methodology employed to address the multiple co-morbidities associated with COPD. They agreed with the TAP's assessment of Scientific Acceptability and raised no new concerns. |
| 3. Usability:<br>3a. Measure performance results are publicly reported: H-9; M-0; L-0; I-0<br>*TAP Discussion:* The TAP believes this subcriterion has been met as NCQA does extensive audits of their material on a regular basis.<br>3b. Measure results are meaningful/useful for public reporting and quality improvement: H-5; M-4; L-0; I-0<br>*TAP Discussion*: The TAP feels the results are usable and understandable.<br>3c. Data and results can be decomposed for transparency and understanding: H-6; M-3; L-0; I-0<br>*TAP Discussion:* The TAP feels this subcriterion has been met as NCQA does extensive audits of their material on a regular basis, and the measure can be deconstructed to facilitate transparency.<br>3d. Harmonized or justification for differences: N/A |
| Overall Usability: H-7; M-7; L-0; I-0<br>Committee Discussion: The Steering Committee valued NCQA's rigorous auditing processes and the transparency with which the developers construct their measures. In addition to being used by health plans, the Committee acknowledged the usefulness of measures for purchasers/providers, giving them much more leverage during negotiations for their annual purchasing agreements. |
| 4. Feasibility:<br>4a. Data elements routinely generated during care process: H-9; M-0; L-0; I-0<br>*TAP Discussion*: The TAP believes this subcriterion has been met as data is a byproduct of care.<br>4b. Data elements available electronically: H-9; M-0; L-0; I-0<br>*TAP Discussion*: The TAP believes this subcriterion has been met; all data is available electronically.<br>4c. Susceptibility to inaccuracies/ unintended consequences identified: H-6; M-3; L-0; I-0<br>*TAP Discussion*: The TAP believes this subcriterion has been met.<br>4d. Data collection strategy can be implemented: H-8; M-1; L-0; I-0<br>*TAP Discussion*: The TAP believes this subcriterion has been met. |
| Overall Feasibility: H-10; M-4; L-0; I-0<br>Committee Discussion: There were no new additional comments from the Steering Committee relating to feasibility of NCQA measures. |

868

| |
|---|
| 1611: ETG Based Pneumonia Cost of Care Measure (Ingenix) |
| Description: The measure focuses on resources used to deliver episodes of care for patients with pneumonia.<br>Pneumonia episodes are defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating pneumonia. A number of resource use measures are defined for pneumonia episodes, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per episode and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons. As requested by NQF, the focus of this submission is for pneumonia episodes and will cover both measures at the pneumonia base and severity level and also a pneumonia composite measure where pneumonia episode results are combined across pneumonia severity levels. At the most detailed level, the measure is defined as the base condition of pneumonia and an assigned level of severity (e.g., resources per episode for pneumonia, severity level 1 episodes). Composite measures can then be created using these measurement units to meet a specific need. For example, a composite measure for pneumonia is derived by combining pneumonia episode results across pneumonia severity levels. Appropriate risk adjustment is applied to support comparisons (e.g., for physician measurement, adjusting for a physician's mix of pneumonia episodes by severity level when supporting a pneumonia composite comparison). The focus of this measure is on pneumonia. However, pneumonia episode results could also be included in a "pulmonary" or other clinical composite for a physician, combining episodes in clinical areas similar to pneumonia. Further, an "overall" composite for a physician can be created, again by aggregating episode results |

across appropriate conditions and severity levels and applying proper risk adjustment when making comparisons.
Resource Use Type: Per episode
Data Type: Administrative claims, Other
Resource Use Service Categories:
Inpatient services: Inpatient facility 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 : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Emergency Medical Services/Ambulance, Home Health, Hospice, Hospital/Acute Care Facility, Imaging Facility, Laboratory, Post Acute/Long Term Care Facility: Nursing Home/Skilled Nursing Facility, Post Acute/Long Term Care Facility: Rehabilitation
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team, Facility, Health Plan, Integrated Delivery System, Population: Community, Population: County or City, Population : National, Population : Regional, Population: State
Measure Developer: Ingenix, 950 Winter Street, Suite 3800, Waltham, Massachusetts, 02451

Committee Recommendation for Endorsement: Y-12; N-4; Abstain-0

Conditions/Questions for Developer:

1. Would it be possible to break down the measure by bacterial versus non-bacterial to try to separate out pneumonia types?

Developer Response:

1. Yes, the measure is stratified. To the extent that administrative claims code the differences in pneumonia types, the measure can be stratified to evaluate resource use differences between pneumonia types.

1.Importance to Measure and Report
1a.High Impact: H-8; M-0; L-0; I-0
*TAP Discussion:* The TAP agreed that pneumonia is a high impact and high cost area.
1b. Resource use/cost problems: H-8; M-0; L-0; I-0
*TAP Discussion:* The TAP believes this subcriterion has been met.
1c. Purpose clearly described: H-8; M-0; L-0; I-0
*TAP Discussion:* The TAP feel the purpose and objective are clear.
1d. Resource use service categories consistent and representative: H-7; M-1; L-0; I-0
*TAP Discussion:* The TAP agrees the service categories are consistent and representative.

Overall Importance: Y-14, N-1
Committee Discussion: The Steering Committee deemed the measure to be important.

2.Scientific Acceptability of Measure Properties:
2a. Overall Reliability: H-3; M-3; L-0; I-1
2a1.Measure well defined and precisely specified: H-3; M-4; L-0; I-0
*TAP Discussion*: Several TAP members were uncomfortable with the lack of transparency in the risk adjustment specifications and felt that the severity weights, particularly for the elderly, were unclear. The panel also had a hard time identifying clean periods. There was a strong feeling that there should be some separation between community-acquired and healthcare-acquired pneumonia, as they represent very different clinical conditions.
2a2. The results are repeatable: H-6; M-1; L-0; I-0
*TAP Discussion:* The TAP had concerns regarding the fact that that there is no way to ascertain how Ingenix came up with the specific weights assigned to comorbidities.
2b. Overall Validity: H-0; M-7; L-0; I-0
2b1. Evidence is consistent with intent: H-4; M-3; L-0; I-0
*TAP Discussion*: The panel again asked for clarification regarding why the measure has different weighted scores for the elderly.
2b2.Score/Analysis: H-0; M-5; L-2; I-0
*TAP Discussion:* The TAP was concerned that they weren't provided with enough information to understand how Ingenix assigned risk scores. Questions regarding how diagnostic descriptions leads to increased utilization were raised. The TAP remained doubtful as to whether this measure should be counted as one distinct population.
2b3. Exclusions: H-2; M-4; L-1; I-0
*TAP Discussion:* The TAP felt that more data around the impact of exclusions (e.g. sensitivity analysis) would be helpful. Ingenix confirmed that there are no clinical exclusions from the measure, only cost exclusions.
2b4. Risk Adjustment: H-1; M-3; L-2; I-1
*TAP Discussion:* The TAP believed that the risk-adjustment methodology is not readily transparent. More information on how risk

scores are assigned was requested from the developers.
2b5. Identification of statistically significant/meaningful differences: H-0; M-7; L-0; I-0
*TAP Discussion:* Data submitted does demonstrate variation in resource use. However, there was a general feeling that meaningfulness is questionable since types of pneumonia cannot be separated out.
2b6. Multiple data sources: N/A (using all administrative data)
2c. Stratification for disparities: H-2; M-5; L-0; I-0
*TAP Discussion:* Gender and age can be stratified, but race data is not available in administrative claims.

Overall Scientifically Acceptable: Yes [Y-13; N-3 (Committee Vote)]
Overall Reliability: H-3; M-11; L-2; I-0
Overall Validity: H-1; M-13; L-2; I-0
Committee Discussion: The Steering Committee agreed that this measure would not be clinically relevant at the physician level due to its limited ability to differentiate between community and hospital acquired pneumonia. In general, the Committee also believed that the "start and stop rules" would be more readily apparent for acute procedure-oriented measures such as knee replacements, as compared with chronic illnesses, which has less clear cut start and stop dates. The Committee reiterated the TAP's concern that Ingenix specified the measure for use in patients over 65 using commercial data to calibrate the model. Commercial patients over 65 are not representative of the general over 65 population.

Usability:
3a. Measure performance results are publicly reported: H-0; M-6; L-1; I-0
*TAP Discussion:* The TAP agrees that despite the fact that multiple care organizations are currently using this measure, the inability to distinguishing between types of pneumonia severely limits the usability of the measure. They concurred that for individual organizations this limitation might be acceptable, but the measure wouldn't be useful as a national consensus standard. . NQF clarified that the measure has a specified for particular levels of analysis, and the ratings need to be reflective of that specification.
3b. Measure results are meaningful/useful for public reporting and quality improvement: H-1; M-5; L-1; I-0
*TAP Discussion:* The TAP agrees that this subcriterion has been met.
3c. Data and results can be decomposed for transparency and understanding: H-1; M-5;L-1; I-0
*TAP Discussion:* The TAP feels the measure would be more transparent if more user-friendly detail were provided.
3d. Harmonized or justification for differences: N/A

Overall Usability: H-3; M-11; L-1; I-1
Committee Discussion: There were no additional concerns identified by the Steering Committee for this criterion.

4. Feasibility:
4a. Data elements routinely generated during care process: H-7; M-0; L-0; I-0
*TAP Discussion*: The TAP believes this subcriterion has been met; data is a byproduct of care.
4b. Data elements available electronically: H-7; M-0; L-0; I-0
*TAP Discussion*: The TAP believes this subcriterion has been met; data available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-1; M-5; L-0; I-1
*TAP Discussion*: The TAP concluded there was a lack of information in the submission regarding data cleaning and missing data to sufficiently understand those areas.
4d. Data collection strategy can be implemented: H-5; M-2; L-0; I-0
*TAP Discussion*: The TAP agrees this subcriterion has been met.

Overall Feasibility: H-1; M-8; L-7; I-0
Committee Discussion: See Ingenix feasibility discussion above.

869

1609: ETG/PEG Based hip/knee replacement Cost of Care Measure (Ingenix)

Description: The measure focuses on resources used to deliver episodes of care for patients who have undergone a Hip/Knee Replacement. Hip Replacement and Knee Replacement episodes are initially defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating the condition. The Procedure Episode Group (PEG) methodology uses the ETG results and further logic to creating a procedure episode that focuses on the Hip Replacement and Knee Replacement component of the care. Procedure episodes identify a unique procedure event as well as the related services performed before and after the procedure including workup and therapy prior to the procedure as well as post-op activities such as repeated surgery and patient follow-up. Together, the ETG and

PEG methodologies identify the services involved in diagnosing, managing and treating patients with Hip/Knee Replacements. A methodology to assign a severity level to each episode is employed to group Hip and Knee Replacement episodes by level of risk.
Resource Use Type: Per episode
Data Type: Administrative claims
Resource Use Service Category: Inpatient services: Inpatient facility 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 : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Emergency Medical Services/Ambulance, Home Health, Hospice, Hospital/Acute Care Facility, Imaging Facility, Laboratory Post Acute/Long Term Care Facility : Nursing Home/Skilled Nursing Facility, Post Acute/Long Term Care Facility : Rehabilitation
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team, Facility, Health Plan, Integrated Delivery System, Population : Community, Population : County or City, Population : National, Population : Regional, Population : State
Measure Developer: Ingenix, 950 Winter Street, Suite 3800, Waltham, Massachusetts, 02451

Committee Recommendation for Endorsement: Y-9; N-7; Abstain-0

Conditions/Questions for Developer: N/A
Developer Response: N/A

1.Importance to Measure and Report –
1a.High Impact: H-6; M-1; L-0; I-0
*TAP Discussion:* The TAP deemed this measure to be a high cost/high impact area.
1b. Resource use/cost problems: H-0; M-2; L-5; I-0
*TAP Discussion:* The TAP felt that the measure would be able to identify large variation in resource use and cost. However, the TAP felt that the developers could have provided more information specifically related to hip/knee replacement variation in resource use in the measure submission.
1c. Purpose clearly described: H-0; M-5; L-1; I-1
*TAP Discussion:* The TAP felt that the purpose was sufficiently described.
1d. Resource use service categories consistent and representative: H-2; M-5; L-0; I-0
*TAP Discussion:* The TAP felt that the resource use service categories were appropriate.

Overall Importance: Y-17, N-0
Committee Discussion: The Steering Committee deemed this measure to be important.

2.Scientific Acceptability of Measure Properties:
2a. Reliability:
2a1.Measure well defined and precisely specified: H-0; M-3; L-4; I-0
*TAP Discussion*: The TAP wanted more information on how the developers handled right and left hip/knee replacement since there is limited ability to distinguish between right/left surgery in the administrative data used. It is important to capture the rate of surgery at the provider level to ensure that the current measure construct does not penalize those providers who chose conservative treatment for low severity patients. The developer should provide more clear information on the clinical logic, including the specific codes that are used to create the episodes. Overall, the TAP wanted more clarity on the clinical construction logic of the episode such as severity level assignments, assignment of claims with two concurrent episodes (i.e. tie breaking logic). The TAP also wanted more information on the procedure definitions, handling of comorbidities and the weighting of multiple co-occurring comorbidities.
2a2. The results are repeatable: H-2; M-5; L-0; I-0
*TAP Discussion:* The TAP wanted additional information on how reliable the physician level scores were over time.
Overall Reliability: H-2; M-4; L-0; I-0
*TAP Discussion:*
2b. Validity
2b1. Evidence is consistent with intent: H-2; M-4; L-1; I-0
TAP Discussion: The TAP felt that the evidence was consistent with the intent of the measure.
2b2.Score/Analysis: H-1; M-4; L-2; I-0
*TAP Discussion:* The TAP discussed the attribution of costs six months before the procedure as too long of a period for a physician based measure. With the current attribution method, it appears to be more appropriate at a plan or system-level rather than an individual provider. These attribution approaches were submitted as guidelines only.
2b3. Exclusions: H-0; M-2; L-4; I-1

*TAP Discussion:* The TAP wanted more information on why low cost outliers were excluded and high cost outliers were windsorized; a sensitivity analysis of this decision was recommended by the TAP. The TAP also recommended that the measure should include a count of high cost outliers if they are going to be winsorized. Information about the high cost outliers might actually drive targeted interventions.

2b4. Risk Adjustment: H-0; M-0; L-6; I-1

*TAP Discussion:* The TAP wanted more information on severity levels on how they related to the risk adjustment model. The TAP agreed that not all of the comorbidities provided in the submission seem appropriate for the population in the measure.

2b5. Identification of statistically significant/meaningful differences:

*TAP Discussion:* There was general agreement that the complexities of the score may make it difficult to discern meaningful differences between providers.

2b6. Multiple data sources: N/A

Overall Validity: H-0; M-1; L-5; I-0

2c. Stratification for disparities: H-1; M-0; L-4; I-2

*TAP Discussion:* Administrative data is limited in its ability to stratify based on race.

Overall Scientifically Acceptable: Yes [Y-11; N-5 (Committee Vote)]

Overall Reliability: H-2; M-14; L-0; I-0

Overall Validity: H-1; M-9; L-6; I-0

Committee Discussion: The Steering Committee was concerned with the lack of specification regarding the measure's use of MSDRG's in the risk-adjustment methodology. Ingenix explained that among the population of patients who undergo knee or hip replacements, there is minimal variation in the underlying co-morbidities. Therefore, the methodology required to adequately risk adjust is much less stringent than it would be if looking at a more complicated condition such as coronary artery disease.

3. Usability:

3a. Measure performance results are publicly reported: H-0; M-5; L-2; I-0

*TAP Discussion:* The TAP was concerned that this ETG was not currently being used as a stand-alone measure and it was unclear if it was currently being publicly reported.

3b. Measure results are meaningful/useful for public reporting and quality improvement: H-0; M-4; L-3; I-0

*TAP Discussion:* The TAP was concerned that this ETG was not currently being used as a stand-alone measure which may impact the need for public reporting.

3c. Data and results can be decomposed for transparency and understanding: H-0; M-3; L-4; I-0

*TAP Discussion:* The TAP expressed concern over the difficulty in understanding the clinical hierarchy and risk model. The lack of clarity in these aspects of the measure makes it difficult to deconstruct the measure for transparency and understanding.

3d. Harmonized or justification for differences: N/A

Overall Usability: H-0; M-12; L-4; I-1

Committee Discussion: The Steering Committee iterated their concern that, because the measure is used as part of a grouper, it is unclear if it is useful as a standalone measure. Additionally, based on the nature of the Ingenix product, hip and knee replacements had been combined into a single measure, which was not believed by some to be the most clinically relevant approach.

4. Feasibility:

4a. Data elements routinely generated during care process: H-5; M-2; L-0; I-0

*TAP Discussion*: The TAP believes this subcriterion has been met; data is a byproduct of care.

4b. Data elements available electronically: H-6; M-1; L-0; I-0

*TAP Discussion*: The TAP believes this subcriterion has been met; data elements that are available electronically.

4c. Susceptibility to inaccuracies/ unintended consequences identified: H-0; M-3; L-4; I-0

*TAP Discussion*: The TAP agrees that much of this surgery is dependent on patient preferences thus the measure should account for these preferences in inclusion and exclusion criteria of the measure. Additionally, providers who treat their patients conservatively can appear to be high cost users since the only patients who get surgery are those who are more severe.

4d. Data collection strategy can be implemented: H-1; M-5; L-1; I-0

TAP Discussion: No additional issues were raised by the TAP.

Overall Feasibility: H-1; M-8; L-7; I-0

Committee Discussion: See Ingenix feasibility discussion above.

## Candidate Consensus Standards Not Recommended for Endorsement

Six candidate consensus standards were not recommended for endorsement because they did not meet NQF criteria; two did not pass scientific acceptability, and the remaining had issues with other criteria.

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 not 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.

***Cardiovascular***



***Non-Condition Specific***



***Bone/Joint***



***Pulmonary***



## Evaluation Summary—Candidate Consensus Standards Not Recommended for Endorsement

| 1591: ETG Based Congestive Heart Failure (CHF) cost of care measure (Ingenix) |
|---|
| Description: The measure focuses on resources used to deliver episodes of care for patients with Congestive Heart Failure (CHF). CHF episodes are defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating CHF. A number of resource use measures are defined for CHF episodes, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per episode and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons. As requested by NQF, the focus of this submission is for CHF episodes and will cover both measures at the CHF base and severity level and also a CHF composite measure where CHF episode results are combined across CHF severity levels. At the most detailed level, the measure is defined as the base condition of CHF and an assigned level of |

severity (e.g., resources per episode for CHF, severity level 1 episodes). Composite measures can then be created using these measurement units to meet a specific need. For example, a composite measure for CHF is derived by combining CHF episode results across CHF severity levels. Appropriate risk adjustment is applied to support comparisons (e.g., for physician measurement, adjusting for a physician's mix of CHF episodes by severity level when supporting a CHF composite comparison). The focus of this measure is on CHF. However, CHF episode results could also be included in a "cardiology", "chronic care", or other clinical composite for a physician, combining episodes in clinical areas similar to CHF. Further, an "overall" composite for a physician can be created, again by aggregating episode results across appropriate conditions and severity levels and applying proper risk adjustment when making comparisons.
Resource Use Type: Per episode
Data Type: Administrative claims, other
Resource Use Service Category: Inpatient services: Inpatient facility 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 : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Emergency Medical Services, Ambulance, Home Health, Hospice, Hospital/Acute Care Facility, Imaging Facility, Laboratory
Level of Analysis: 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
Measure Developer: Ingenix, 950 Winter Street, Suite 3800, Waltham, Massachusetts, 02451

Committee Recommendation for Endorsement: Y-6; N-8; Abstain–0 (re-vote) [Y-10; N-8; Abstain-0 (initial vote)]

Conditions/Questions for Developer:

1. Why are some of the codes, typically seen in congestive heart failure measures, excluded?
2. How are hospitalizations that occur during the course of the measure handled?
3. Does the episode include events that occur before and/or after the episode?

Developer Response:

1. Ingenix excluded the codes that were specific to diastolic heart failure (as this is a systolic and diastolic/systolic mix measure); if those codes were included it would have created another episode. Ingenix includes codes that were both systolic and diastolic, and used them as a marker to increase the severity score for the episode.
2. Hospital admissions that occurred during the course of the measure that are coded for congestive heart failure are included in the measure; hospitalizations are not used for severity adjustment. If the hospital admission date occurs during the measurement year, then the admission is included in that measurement year.
3. No, this measure is insulated from events that occur before or after the episode.

1. Importance to Measure and Report

1a.High Impact: H -8; M-0; L-0; I-0
*TAP Discussion:* The TAP believes this is a high impact, high cost area that is important to measure and report.
1b. Resource use/cost problems: H -8; M-0; L-0; I-0
*TAP Discussion:* The TAP believes this subcriterion has been met.
1c. Purpose clearly described: H -5; M-3; L-0; I-0
*TAP Discussion:* The TAP believes the purpose of the measure is clearly described.
1d. Resource use service categories consistent and representative: H -7; M-1; L-0;I-0
*TAP Discussion:* The TAP believes this subcriterion has been met; the resource use service categories are consistent and representative of the measure.

Overall Importance: Yes [Y-17; N-1 (Committee Vote)]
Committee Discussion: The Steering Committee believes this is a high impact, high cost area and that the measure has been clearly described. This criterion has been met.

2. Scientific Acceptability of Measure Properties:
2a. Reliability:
2a1. Well defined/precise specifications: H -3; M-4; L-0; I-1
*TAP Discussion:* The TAP believed there was a bit of confusion around the term, "congestive heart failure", it was brought up that not all "heart failure" is necessarily "congestive" and there needs to be more clarification around the use of this term. The TAP agrees that this measure is targeting systolic heart failure and then a mix of systolic/diastolic heart failure. Ingenix also has a diastolic heart failure measure, but it has not been submitted for NQF endorsement. When the ICD9 code exists for systolic and diastolic – it's a marker for severity adjustment. Overall, the TAP believes that the clinical and construction logic of the measure was described in sufficient detail and users will be able to implement the measure as described.
2a2. Reliability testing: H -7; M-1;L-0;I-0
*TAP Discussion:* The TAP believes this measure has demonstrated extensive benchmarking and comparisons; however they would have liked to see more external comparisons. The testing data submitted was from nine health care organizations, all large commercial insurers that vary geographically. Ingenix demonstrated reliability by performing parallel development of the data by using two independent approaches. These two different approaches led to the same results as levels near 99.9%The data was tested primarily on commercial databases, however some Part C plan Medicare patients were also included. It is important to note that this measure was submitted for use in the commercial, less than 65 years old population.

2b. Validity:
2b1. Specifications consistent with resource use/cost problem: H -2;M-2; L-0; I-0
*TAP Discussion:* The TAP agrees that the specifications are consistent with the resource use.
2b2. Validity testing: H -4; M-4; L-0; I-0
*TAP Discussion:* The TAP believes Ingenix has sufficiently demonstrated face validity.
2b3. Exclusions: H -4; M-3; L-1; I-0
*TAP Discussion:* There are no exclusions within this measures, the TAP believes this subcriterion has been met.
2b4. Risk adjustment: H -4; M-2; L-0; I-1
*TAP Discussion:* The TAP believes that this risk adjustment appears to be somewhat circular – the measure is risk adjusted if the individual was hospitalized during the year – if the provider is using a large amount of resources, inevitably there will be more diagnoses in that measurement period, which would in turn also affect severity level category. Ingenix has made it clear that they are not using utilization to directly risk-adjust the cost of the episode. There is a lack of information in terms of the variables selected for inclusion in the calibration of the risk model, the risk groups selected in terms of a cutoff for the severity score, and there is no rationale presented for why this cutoff point has been chosen.
2b5. Identification of statistically significant/meaningful differences: H -2; M-1; L-3; I-1
*TAP Discussion*: The TAP believes there is little information to compare statistical versus practical significance for this measure. The measure allows the user to determine what is clinically significant based on confidence intervals. The sample size appears sufficient enough to obtain a confidence interval that it will be useful to establish differences that are clinically and statistically significant. Ingenix has created confidence intervals around the observed to expected ratio The minimum sample size to detect statistically significant differences depends upon the case mix of the providers and the variation in performance across providers..
2b6. Multiple data sources: N/A
2c. Stratification for disparities: H-0; M-0; L-0; I-0; N/A-8
*TAP Discussion:* Due to the limitations in the administrative claims data, at this time the measure does not stratify for disparities.

Overall Reliability: H-3, M-12, L-2, I-0
Overall Validity: H-1, M-13,L-4, I-0
Overall Scientific Acceptability: Yes [Y-14; N-4(Committee Vote)]
Committee Discussion: The Steering Committee discussion focused on how clearly specified the codes used with the measure are, and how well they capture systolic heart failure. This is a measure of systolic heart failure, a paired measure of diastolic heart failure from Ingenix exists but they did not submit it to the project. Because the Steering Committee could not take into account the existence of the diastolic measure, there was concern around the completeness and accuracy with which this measure would capture systolic heart failure. The diagnosis codes specified are limited to the 428 codes that used the word "systolic", they do not use some of the 404s and 402s that the other measures have used to capture the larger heart failure population. The measure specifications have been in use for a significant amount of time; Ingenix has demonstrated that if this measure is used in the same population, at the same time, then the result will be the same roughly 99.9% of the time. The Steering Committee discussed how there are carve outs for mental health & pharmacy data and therefore comparisons within the health plan are the same or likely to be the same. However, when comparing

across health plans or across physician groups validity may become an issue when there are differences in the completeness of the data submitted. The Steering Committee expressed concerns over the reliability, validity and risk adjustment method. Specifically, that the measure may be adjusting for comorbidities identified during the measurement period as opposed to comorbidities identified prior to the episode. There was also concern that the risk adjustment may be "over –adjusting", or possibly "adjusting away" significant differences.

3. Usability:
3a. Measure performance results are publicly reported: H-1; M-1; L-2;I-2
*TAP Discussion:* The TAP was concerned with the availability of this data to the public and requested clarification from NQF on what is required for "public reporting". The measures are widely used by providers to compare to one another. The results of this measure also allow for provider profiling, provider report cards and there is a cost base analysis for the members to estimate what the cost of the service would be, including the out of pocket expense. Since this measure is reported within a suite of measures, it has not been broken out individually for reporting or use in quality improvement.
3b. Measure results are meaningful/useful for public reporting and quality improvement: H-3; M-1; L-0; I-2
*TAP Discussion:* The TAP agrees that more information would be needed to explain the results of this measure to the public and to be used for internal quality improvement.
3c. Data and results can be decomposed for transparency and understanding: H-0; M-2; L-3;I-1
*TAP Discussion:* The TAP agrees there are challenges for the use of this measure, which include its complexity and lack of clarity in the specifications. TAP also agrees it is difficult to assess the extent to which the measure can be decomposed as it is currently specified.
3d. Harmonized or justification for differences: N/A

Overall Usability: H-0; M-10; L-7; I-0, N/A-0
Committee Discussion: The Steering Committee discussed the fact that more information would be needed to explain the results of this measure to the public and to be used for internal quality improvement. The Steering Committee believes there are challenges for the use of this measure, which include its complexity and lack of clarity in the specifications. The Steering Committee agrees it is difficult to assess the extent of which the measure can be decomposed as it is currently specified.

4. Feasibility:
4a. Data elements routinely generated during care process: H-5; M-0; L-0; I-1
*TAP Discussion:* The TAP believes that this sub criterion has been met; all of the data elements are generated during the care process.
4b. Data elements available electronically: H-5;M-0; L-0;I-1
*TAP Discussion:* The TAP believes that this sub criterion has been met; all of the data is available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-0; M-4; L-1; I-1
*TAP Discussion:* The TAP noted that Ingenix does not have a formal audit system to ensure that all of the numbers are included & correct. In general, when dealing with any measure that uses administrative data there are various inaccuracies, pertaining particularly to coding inaccuracies and variation.
4d. Data collection strategy can be implemented: H-3; M-0; L-1; I-2
*TAP Discussion:* The majority of the TAP agreed that barriers to use are minimal. (NQF Note: This is prior to the submission of product pricing information shared only with the Steering Committee)

Overall Feasibility: H-2; M-8; L-7; I-1
Committee Discussion: See Ingenix feasibility discussion above.

895

1594 ETG Based Coronary Artery Disease (CAD) cost of care measure (Ingenix)

Description: The measure focuses on resources used to deliver episodes of care for patients with CAD. CAD episodes are defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating CAD. A number of resource use measures are defined for CAD episodes, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per episode and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons. As requested by NQF, the focus of this submission is for CAD episodes and will cover both measures at the CAD base and severity level and also a CAD composite measure where CAD episode results are combined across CAD severity levels. At the most detailed level, the measure is defined as the base condition of CAD and an assigned level of severity (e.g., resources per episode for CAD, severity level 1 episodes). Composite measures can then be created using these measurement units to meet a specific need. For example, a composite measure for CAD is derived by combining CAD episode results across CAD

severity levels. Appropriate risk adjustment is applied to support comparisons (e.g., for physician measurement, adjusting for a physician's mix of CAD episodes by severity level when supporting a CAD composite comparison). The focus of this measure is on CAD. However, CAD episode results could also be included in a "cardiology", "chronic care", or other clinical composite for a physician, combining episodes in clinical areas similar to CAD. Further, an "overall" composite for a physician can be created, again by aggregating episode results across appropriate conditions and severity levels and applying proper risk adjustment when making comparisons.
Resource Use Type: Per episode
Data Type: Administrative claims, other
Resource Use Service Category: Inpatient services: Inpatient facility 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 : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Emergency Medical Services, Ambulance, Home Health, Hospice, Hospital/Acute Care Facility, Imaging Facility Laboratory
Level of Analysis: 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
Measure Developer: Ingenix, 950 Winter Street, Waltham, Massachusetts, 02154

Committee Recommendation for Endorsement: Y-5; N-9; Abstain – 0 (re-vote) [Y-8; N-10; Abstain-0 (initial vote)]

1. Importance to Measure and Report
1a.High Impact: H-5; M-0; L-0; I-0
*TAP Discussion:* The TAP believes this is a high impact, high cost area; this sub criterion has been met.
1b. Resource use/cost problems: H-5; M-0; L-0; I-0
*TAP Discussion:* The TAP believes this subcriterion has been met.
1c. Purpose clearly described: H-5' M-0; L-0; I-0
*TAP Discussion:* The TAP believes this subcriterion has been met; the measure purpose is clearly described.
1d. Resource use service categories consistent and representative: H-3; M-2; L-0; I-0
*TAP Discussion:* The TAP believes this subcriterion has been met; the resource use categories are consistent and representative.

Overall Importance: Y-16, N-1 (Committee Vote)
Committee Discussion: The Steering Committee believes this is a high impact, high cost area and that the measure has been clearly described. This criterion has been met.

2. Scientific Acceptability of Measure Properties:
2a. Reliability:
2a1. Well defined/precise specifications: H-3; M-1; L-0; I-0
*TAP Discussion:* The diagnoses codes for this measure are the 410s through 414s and then the 429s, all of which represent complications of myocardial infarction. These codes seem comprehensive for identifying patients with coronary artery disease; however, the Steering Committee raised the question if the populations are similar enough that the user can reasonably make inferences about the resource use needed for each type of cardiac episode. Overall, the measure is very well specified and is being used across different health plans.
2a2. Reliability testing: H-3; M-1; L-0; I-0
*TAP Discussion:* The measure is specified in a way that it has been used over a long period of time, Ingenix demonstrated that if the user uses the same measure in the same population then the result will be the same. The TAP believes this subcriterion has been met.
2b. Validity:
2b1. Specifications consistent with resource use/cost problem: H-3; M-1; L-0;I-0
*TAP Discussion:* The TAP believes this subcriterion has been met; a specific population is defined and measured.
2b2. Validity testing: H-3;M-0; L-0; I-0
*TAP Discussion:* The TAP believes Ingenix has sufficiently demonstrated face validity.
2b3. Exclusions: H-2;M-1; L-0; I-0
*TAP Discussion:* There are no exclusions within this measures, the TAP believes this subcriterion has been met.
2b4. Risk adjustment: H-2; M-1; L-0;I-0
*TAP Discussion:* The TAP requested that the developer demonstrate proof of the concept that this is accurately accounting for differences in the population – the risk adjustment method does not appear to be robust. Additional information the model's goodness of fit was requested. NQF staff is working with Ingenix to supply this information to the Steering Committee.
2b5. Identification of statistically significant/meaningful differences: H-1;M-0; L-1;I-1
*TAP Discussion:* The Steering Committee believes that this measure did not identify statistically significant or meaningful differences across groups. There was general concern that something may be classified as statistically significant, when it is not clinically significant.
2b6. Multiple data sources: N/A
*TAP Discussion:* N/A
2c. Stratification for disparities: H-0; M-0; L-0; I-0; N/A-8
*TAP Discussion:* Due to the limitations in the administrative claims data, at this time the measure does not stratify for disparities.

Overall Reliability : H-5; M-11; L-2; I-0
Overall Validity: H-2;M-10; L-6;I-0
Overall Scientifically Acceptable: Yes [Y-12; N- 5 (Committee vote)]
Committee Discussion: The Steering Committee agreed that the measure accurately identified the primary incurring diagnoses codes as 410s through 414s. Within those strata there is a range of conditions – ranging from chronic, stable coronary artery disease to patients with cardiogenic shock complicated by a flail mitral posterior leaflet. The Steering Committee discussed how there is a large spectrum of risk adverse outcomes within this population. Furthermore, this carries the risk of different resource use for each specific condition included in the measure. The measure was submitted for implementation across various levels of analysis, however for individual clinicians there is not a sample size guideline. Regarding specific reliability testing, the measure is specified in a way that it has been used over a long period of time. The Steering Committee discussed how there are carve outs for mental health & pharmacy data and therefore comparisons within the health plan are the same or likely to be the same. However, when comparing across health plans or across physician groups validity may become an issue. There were concerns around the risk adjustment method. Specifically, the Committee was concerned that the measure may be adjusting for comorbidities identified during the measurement episode as opposed to comorbidities identified prior to the episode. There was also concern that the risk adjustment may be "over –adjusting", or possibly "adjusting away" significant differences.

3. Usability:
3a. Measure performance results are publicly reported: H-0;M-1;L-1; I-1
*TAP Discussion:* The TAP was concerned with the availability of this data to the public and requested clarification from NQF on what is required for "public reporting". The measures are widely used by providers to compare to one another. The results of this measure also allow for provider profiling, provider report cards and there is a cost base analysis for the members to estimate what the cost of the service would be, including the out of pocket expense. Since this measure is reported within a suite of measures, it has not been broken out individually for reporting or use in quality improvement.
3b. Measure results are meaningful/useful for public reporting and quality improvement: H-0; M-2; L-1;I-0

*TAP Discussion:* The TAP agrees that more information would be needed to explain the results of this measure to the public and to be used for internal quality improvement.
3c. Data and results can be decomposed for transparency and understanding: H-0; M-3;L-0;I-0
*TAP Discussion:* The TAP agreed there are challenges for the use of this measure, which include its complexity and lack of clarity in the specifications. TAP also agreed it is difficult to assess the extent of which the measure can be deconstructed for understanding as it is currently specified.
3d. Harmonized or justification for differences: N/A

Overall Usability: H-1; M-11; L-4; I-1
Committee Discussion: The Steering Committee agrees that more information would be needed to explain the results of this measure to the public and to be used for internal quality improvement. The Steering Committee discussed the challenges for the use of this measure, which include its complexity and lack of clarity in the specifications. The Steering Committee agrees it is difficult to assess the extent of which the measure can be decomposed as it is currently specified.

4. Feasibility:
4a. Data elements routinely generated during care process: H-3; M-0; L-0; I-0
*TAP Discussion:* The TAP believes that this sub criterion has been met; all of the data elements are generated during the care process.
4b. Data elements available electronically: H-3; M-0; L-0;I-0
*TAP Discussion:* The TAP believes that this sub criterion has been met; all of the data is available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-2; M-1; L-0; I-0
*TAP Discussion:* The TAP noted that Ingenix does not have a formal audit system to ensure that all of the numbers are included & correct. In general, when dealing with any measure that uses administrative data there are various inaccuracies, pertaining particularly to coding inaccuracies and variation.
4d. Data collection strategy can be implemented: H-2;M-0; L-1; I-0
*TAP Discussion:* The majority of the TAP agreed that barriers to use are minimal. (NQF Note: This is prior to the submission of product pricing information shared only with the Steering Committee)

Overall Feasibility: H-3; M-8; L-6; I-1
Committee Discussion: See Ingenix feasibility discussion above.

896

1599: ETG Based Non-Condition Specific cost of care measure (Ingenix)

Description: The measure focuses on resources used to diagnose, manage and treat a population of patients (non-condition specific) during a defined 12-month period of time. The population included in the measurement can be described generally. Examples include a population of individuals enrolled with a health plan, individuals assigned to a patient-centered medical home or accountable care organization (ACO), or a panel of individuals managed by a primary care physician (PCP). A number of resource use measures are defined for this measure set, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per member per month and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons. Risk adjustment is based on the measure of risk assigned to each individual using the Episode Risk Group (ERG) methodology.
Resource Use Type: Per capita (population- or patient-based)
Data Type: Administrative claims
Resource Use Service Category: Inpatient services: Inpatient facility 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 : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Emergency Medical Services, Ambulance, Home Health, Hospice, Hospital/Acute Care Facility, Imaging Facility, Laboratory
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team, Facility, Health Plan, Integrated Delivery System Population : County or City, Population : National, Population : Regional, Population : states
Measure Developer: Ingenix, 950 Winter Street, Suite 3800, Waltham, Massachusetts, 02451

Committee Recommendation for Endorsement: Y-5; N-9; Abstain-0 (re-vote) [ Y-12; N-6; Abstain-0 (initial vote)]

Conditions/Questions for Developer:
1. How does the risk score correlate with the actual expenditures?

2. What is the distinction between ETGs and ERGs?
3. Can this measure be applied to the Medicare population?
4. Have there been any changes in the underlying risk model used in the ETGs since what has been published on the Ingenix web site a year ago?
5. How are the carve outs, pharmacy and mental health data handled? How was this data validated?

Developer Response:

1. Ingenix provides options for expenditure thresholds for a patient's annual member costs: $25,000, $100,000, and $250,000. Ingenix explained that these thresholds would vary depending on the application.
2. ETGs are episode-based measures. For example, an episode of diabetes, congestive heart failure or COPD--the severity models are built separately for each of the conditions which allows for risk adjustment for each separate condition-based episode. The results are then tagged for each episode for a member not only by condition, but also by the level of severity. There are hundreds of ETGs that map into the ERGs. Ingenix maps to the ERG designed for the population-based risk adjustment; they weight each of the ERG markers to the final ERG score. The ERGs looks at age, in which case they may be applied to the Medicare population, however not all of the ETGs take age into account in the risk adjustment model. During the developer testing they didn't find that age had much explanatory power so they are not included in all of the ERGs. The ERG will point to a different weight depending on the age of the individual. However, since this measure has only been tested in a commercial database, per NQF policy, it can only be endorsed for use in commercial populations.
3. The ETG models and the risk models related to the ETGs have not been updated or recalibrated within the last year; therefore the information on the Ingenix website is still applicable.
4. Ingenix works with a population that has pharmacy and medical data. Mental health is excluded because the claims are not often available in addition to lack of coding for mental health services. Pharmacy data hasn't been an issue because it's up to the user whether they want to include and compare populations who have pharmacy data. The methodology can be adjusted, you are able to a have a mixed population of both medical and pharmacy benefits, and the user is able to isolate the medical resource use data if they choose to.

1. Importance to Measure and Report :Y-16; N-0

*Committee Discussion:* This criterion was also discussed during the June 6 conference call. To access the summary of this call, **click here.**

1a.High Impact: H-15; M-1; L-0; I-0

*Committee Discussion:* The Steering Committee has deemed the measure focus to be high impact.

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

*Committee Discussion:* The Steering Committee agrees this criterion has been met.

1c. Purpose clearly described: H-12; M-4; L-0; I-0

*Committee Discussion:* The Steering Committee believes the measure has met this sub criterion, as the measure's purpose is clearly described.

1d. Resource use service categories consistent and representative: H-8; M-8; L-0; I-0

*Committee Discussion:* The resource use service categories are representative of the measure intent and focus.

2. Scientific Acceptability of Measure Properties: Yes [Y-9; N-6 (Committee Vote)]

2a. Overall Reliability: H-8; M-7; L-1; I-0

*Committee Discussion:* The Ingenix team has a robust system where they double code the data – the steps that lead to the production of the data has a 99.9% match between the two approaches.. The Committee agreed that tables present measure results it is unclear if they actually represent that the measure is reliable.

2a1.Measure well defined and precisely specified: H-10; M-5; L-1; I-0

*Committee Discussion:* This measure appears to be well defined and specified. This methodology is used in a number of organizations and appears to work well. This sub criterion has been met.

2a2. Reliability Testing: H-9; M-7; L-0; I-0

*Committee Discussion:* The Committee agreed that this sub criterion has been met; the results have shown to be repeatable. The Committee suggested more robust reliability testing methods should be explored.

2b. Overall Validity: H-2; M-10; L-3; I-0

*Committee Discussion*: In the submission, Ingenix states that they apply the methodology to data from several different organizations, but this is not detailed in any of the results. Face validity was tested however there is not any description of the results within the submission. The tables that were submitted to demonstrate validity are not clearly labeled or defined.

2b1. Specifications consistent with intent: H-7; M-8; L-1; I-0

*Committee Discussion:* The Committee agrees the specifications are consistent with the intent.
2b2.Validity Testing: H-0; M-8; L-6; I-0
*Committee Discussion:* This measure has been demonstrated to meet the requirement for face validity.
2b3. Exclusions: H-9; M-4; L-2 ; I-0
*Committee Discussion:* There are no exclusions based on cost or other criteria. The Committee reiterated concerns with comparability for plans that have pharmacy carve outs or do not have pharmacy data to those that do.
2b4. Risk Adjustment: H-6; M-8; L-1; I-0
*Committee Discussion:* When looking at the ETG codes, a severity score is assigned; the methodology then takes into account the ETG severity score and the number of comorbidities. A retrospective model contains the observed episodes that may occur during that year, but a user will not be able to observe any markers or costs for people who did not undergo services. The ERG risk level determines the individual's ERG risk score which drives the risk adjustment. The Committee acknowledged this methodology is very complex and not completely understood by all members.
2b5. Identification of statistically significant/meaningful differences: H-5 ; M-7 ; L-3; I-0
*Committee Discussion:* There is a way to stratify those with or without pharmacy data. The Committee expressed concern that valid comparisons cannot be made across organizations with different levels of data completeness and consistency.
2b6. Multiple data sources: N/A
2c. Stratification for disparities: H-0; M-4; L-2; I-9
*Committee Discussion:* This measure does not stratify by race and ethnicity. This may be possible in the future, but at the present time this information is not available.

3. Overall Usability: H-0; M-10; L-5; I-0
*Committee Discussion:* The Committee questioned on whether this measure has been featured in peer reviewed articles; the developer was unaware of any that could be shared with the Committee. The developers explained that this measure is currently being used to profile physicians. They are unaware of any efforts to publicly report the results, even within health plans to their covered lives.
3a. Measure performance results are publicly reported: H-0; M-4; L-6; I-4
*Committee Discussion:* Ingenix conducted a survey of their customers, some users are publicly reporting the data and others are sharing information with physicians for incentive based programs. Some users have decided to put the information on a website that goes to their providers, which allows them to access their risk scores and score card. Providers are then able to drill down on the scorecard to the claim base level, the patient level and then the overall claims level.
3b. Measure results are meaningful/useful for public reporting and performance improvement:
*Committee Discussion*: H-3; M-6; L-3; I-3
3c. Data and results can be decomposed for transparency and understanding: H-1; M-8; L-5; I-1
*Committee Discussion:* While Ingenix has a transparency website open to the public which explains the methodology and approach to measuring resources, the submission reviewed by the Committee was admittedly complex and at times difficult to identify the relevant information.
3d. Harmonized or justification for differences: N/A

4. Feasibility: H-3; M-8, L-6, I-0
4a. Data elements routinely generated during care process: H-13; M-2; L-2; I-0
*Discussion:* The Steering Committee believes that this sub criterion has been met; all of the data elements are generated during the care process.
4b. Data elements available electronically: H-14, M-4, L-0, I-0
*Discussion:* The Steering Committee believes that this sub criterion has been met; all of the data is available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-5, M-9, L-3; I-0
*Discussion:* Mental health is not available and pharmacy data rarely is, when pharmacy data is included it is stratified. Ingenix does not have a formal audit system to ensure that all of the numbers are included & correct. In general, when dealing with any measure that uses administrative data there are various inaccuracies, pertaining particularly to coding inaccuracies and variation. Ingenix provides guidelines how to use small volumes/ sample sizes, however there is not content available to demonstrate this approach. This measure appears less prone to "gaming", as there is not much a user can do to manipulate the start or end of an episode.
4d. Data collection strategy can be implemented: H-1, M-10, L-13, I-1
*Discussion:* See Ingenix feasibility discussion above.

897

898

1603: ETG/ PEG Based Hip Fracture Cost of Care measure (Ingenix)

Description: The measure focuses on resources used to deliver episodes of care for patients with Hip Fracture. Hip Fracture episodes are defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating Hip Fracture. A number of resource use measures are defined for Hip Fracture episodes, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per episode and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons. As requested by NQF, the focus of this submission is for Hip Fracture episodes and will cover both measures at the Hip Fracture base and severity level and also a Hip Fracture composite measure where Hip Fracture episode results are combined across Hip Fracture severity levels. At the most detailed level, the measure is defined as the base condition of Hip Fracture and an assigned level of severity (e.g., resources per episode for Hip Fracture, severity level 1 episodes). Composite measures can then be created using these measurement units to meet a specific need. For example, a composite measure for Hip Fracture is derived by combining Hip Fracture episode results across Hip Fracture severity levels. Appropriate risk adjustment is applied to support comparisons (e.g., for physician measurement, adjusting for a physician's mix of Hip Fracture episodes by severity level when supporting a Hip Fracture composite comparison).The focus of this measure is on Hip Fracture. However, Hip Fracture episode results could also be included in an "orthopedics", "acute care", or other clinical composite for a physician, combining episodes in clinical areas similar to Hip Fracture. Further, an "overall" composite for a physician can be created, again by aggregating episode results across appropriate conditions and severity levels and applying proper risk adjustment when making comparisons.
Resource Use Type: Per episode
Data Type: Administrative claims, Other
Resource Use Service Categories:
Inpatient services: Inpatient facility services; Admissions/discharged; Ambulatory services: Outpatient facility services; Emergency Department; Pharmacy; Evaluation and management; Procedures and surgeries; Imaging and diagnostic; Lab services
Care Setting: Ambulatory Care : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Emergency Medical Services/Ambulance, Home Health, Hospice, Hospital/Acute Care Facility, Imaging Facility, Laboratory, Post Acute/Long Term Care Facility: Nursing Home/Skilled Nursing Facility, Post Acute/Long Term Care Facility: Rehabilitation
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team, Facility, Health Plan, Integrated Delivery System, Population: Community, Population: County or City, Population : National, Population : Regional, Population: State
Measure Developer: Ingenix, 950 Winter Street, Suite 3800, Waltham, Massachusetts, 02451

Committee Recommendation for Endorsement: This measure did not pass the scientific acceptability criterion, and is not recommended for endorsement.

Conditions/Questions for Developer:

1. Why are different age groups assigned the same risk coefficients, when they will have extremely different risk factors?
2. How does the episode grouper work in terms of low and high outliers? Are you able to provide information on exactly how many episodes have been excluded?
3. Why do you cut the low cost episodes from being included in the measure?

Developer Response:

1. This represents a limitation of the data set. Due to the minimal number of people over 65 in commercial programs, we didn't have the numbers to further stratify.
2. We exclude cases that are low in cost. We have the data to talk about the number of cases that are excluded by varying a low outlier, yes.
3. The hypothesis that that these low cost episodes – ones under 2.5 percent – are either mistakes or miscodes. They are probably incomplete episodes, so we don't count them.

1. Importance to Measure and Report

1a.High Impact: H-2; M-1; L-2; I- 0
*TAP Discussion:* There was general agreement that hip fracture is a major cause of morbidity, mortality and high resource use. The TAP did, however, question the importance of measuring hip fractures in a predominately under 65 group of patients. Ingenix acknowledged that this was a significant limitation of using administrative data.
1b. Resource use/cost problems: H-2; M-2; L-1; I-0
*TAP Discussion:* No issues were identified.
1c. Purpose clearly described: H-1; M-4; L-0; I-0

*TAP Discussion:* No issues were identified.
1d. Resource use service categories consistent and representative: H-2; M-2; L-1; I-0
*TAP Discussion:* The TAP were concerned that resource use service categories omit nursing homes and inpatient or outpatient rehab services.

Overall Importance: Y-10, N-6
Committee Discussion: The Committee agreed that hip fractures are a high impact area of healthcare. They were concerned, however, that the measure did not include populations of patients over 65, where the vast majority of hip fractures would occur, and where the nature of hip fractures is a significantly different than it is for younger populations. Ingenix reminded the Committee that the measure was tested in a commercial database, not a Medicare database, and would therefore be endorsed as such. The Committee ultimately questioned whether it was important to measure hip fractures in a younger population at all.

2. Scientific Acceptability of Measure Properties:
2a. Overall Reliability: H-1; M-0; L-4; I-0
2a1.Measure well defined and precisely specified: H-1; M-2; L-2; I-0
*TAP Discussion*: The TAP was concerned that the measure didn't capture certain co-morbid conditions such as dementia which are critical to understanding resource use for this clinical condition. There was substantial unease that the data does not examine the Medicare population, where the majority of hip-fractures occur.
2a2. The results are repeatable: H-1; M-2; L-2; I-0
*TAP Discussion:* The panel questioned whether one could infer grouper reliability from the tables submitted by Ingenix. Ingenix explained that the tables illustrate expected variability in results and point to a relatively consistent cost across health care organizations.
2b. Overall Validity: H-0; M-1; L-3; I-0
2b1. Evidence is consistent with intent: H-0; M-0; L-5; I-0
*TAP* Discussion: The TAP reiterated their concern that the measure hasn't captured the patient population most likely to be affected by hip fractures. Therefore, the measure may have limited applicability, due to the limitations of using only commercial data. The panel also felt that hip fractures in younger populations versus older populations represent two very different clinical situations.
2b2.Score/Analysis: H-0; M-1; L-4; I-0
*TAP Discussion:* The TAP was uncomfortable with the fact that all age groups were assigned the same risk coefficients. Ingenix explained that this also represents a limitation of the data set, where they did not have the numbers over 65 to further stratify. Members of the panel believed that certain clinically relevant co-morbidities and complications such as dementia and post-op delirium should be reported on in a hip-fracture measure.
2b3. Exclusions: H-0; M-1; L-4; I-0
*TAP Discussion:* The TAP felt that the reasoning behind the exclusion criteria was unclear and not based on clinical evidence.
2b4. Risk Adjustment: H-0; M-0; L-4; I-1
*TAP Discussion:* The developer described how the measure contains low dollar exclusions. The assumption is that these claims represent incomplete episodes.
2b5. Identification of statistically significant/meaningful differences: H-0; M-0; L-4; I-1
*TAP Discussion:* There was a discussion regarding the relative cost of care ratio and a question about what numbers represent statistically significant differences. Ingenix explained that the numbers would depend on the confidence interval, the underlying variance of episode cost and the number of total cases.
2b6. Multiple data sources: N/A (using all administrative data)
2c. Stratification for disparities: H-0; M-1; L-1; I-3
*TAP Discussion:* Racial disparities were addressed in the submission, but the data limits a further examination into these disparities.

Overall Scientifically Acceptable: No [Y-7; N-10 (Committee Vote)]
Overall Reliability: H-1; M-11; L-3; I-2
Overall Validity: H-0; M-6; L-10; I-0
Committee Discussion: The Committee believed the measure was limited in its clinical construction logic as a result of its reliance upon commercial data, where the population of patients with hip fractures was notably low. Thus, the testing completed by Ingenix for this measure represented a fairly *uncommon* condition – hip fractures in under 65's – when the majority of hip fractures are much more common and different clinically. The Committee agreed, therefore, that significant and meaningful differences could not be produced by this measure, particularly when reporting at an individual physician level. Furthermore, the Committee were concerned with the fact that the grouper function was not tested or reported on, and Ingenix provided no information comparing scoring of attribution over episodes of time

Usability:

3a. Measure performance results are publicly reported: H-0; M-2; L-3; I-0
*TAP Discussion:* The TAP believes this subcriterion has been met.
3b. Measure results are meaningful/useful for public reporting and quality improvement: H-0; M-1; L-4; I-0
*TAP Discussion:* The TAP acknowledged the impressive amount of work Ingenix put into this measure, but again articulated concern that the measure would have limited meaningful use as it is not capturing the appropriate population. The panel was uneasy with the grouping of two clinically different age cohorts together into one measure; they felt that the clinical situation, treatment path and mortality for a younger population with hip fractures versus an older population were different enough to warrant two separate measures.
3c. Data and results can be decomposed for transparency and understanding: H-0; M-2; L-3; I-0
*TAP Discussion:* The TAP agrees this subcriterion has been met.
3d. Harmonized or justification for differences: N/A

Overall Usability: This measure did not pass the scientific acceptability criterion. As a result, the Committee did not discuss usability.

4. Feasibility:
4a. Data elements routinely generated during care process: H-3; M-1; L-1; I-0
*TAP Discussion:* The TAP agrees that this subcriterion has been met; all data is routinely generated through the care process.
4b. Data elements available electronically: H-4; M-0; L-1; I-0
*TAP Discussion:* The TAP agrees that this subcriterion has been met; all data is available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-1; M-1; L-3; I-0
*TAP Discussion:* The TAP believe that this subcriterion has been met, however Ingenix does not have a formal audit system in order to monitor for inaccuracies.
4d. Data collection strategy can be implemented: H-0; M-2; L-2; I-1
*TAP Discussion:* The TAP believe that this subcriterion has been met. (NQF Staff Note: this is prior to the submission of product pricing information reviewed by the Steering Committee only.)

Overall Feasibility: This measure did not pass the scientific acceptability criterion. As a result, the Committee did not vote on feasibility.

899

1605: ETG Based Asthma Cost of Care Measure(Ingenix)

Description: The measure focuses on resources used to deliver episodes of care for patients with Asthma.
Asthma episodes are defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating asthma. A number of resource use measures are defined for asthma episodes, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per episode and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons.
As requested by NQF, the focus of this submission is for Asthma episodes and will cover both measures at the Asthma base and severity level and also an Asthma composite measure where Asthma episode results are combined across Asthma severity levels. At the most detailed level, the measure is defined as the base condition of Asthma and an assigned level of severity (e.g., resources per episode for Asthma, severity level 1 episodes). Composite measures can then be created using these measurement units to meet a specific need. For example, a composite measure for Asthma is derived by combining Asthma episode results across Asthma severity levels. Appropriate risk adjustment is applied to support comparisons (e.g., for physician measurement, adjusting for a physician's mix of Asthma episodes by severity level when supporting an Asthma composite comparison). The focus of this measure is on Asthma. However, Asthma episode results could also be included in a "pulmonologist", "chronic care", or other clinical composite for a physician, combining episodes in clinical areas similar to Asthma. Further, an "overall" composite for a physician can be created, again by aggregating episode results across appropriate conditions and severity levels and applying proper risk adjustment when making comparisons.
Resource Use Type: Per episode
Data Type: Administrative claims, Other
Resource Use Service Categories:
Inpatient services: Inpatient facility 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 : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Emergency Medical Services/Ambulance, Home Health, Hospice, Hospital/Acute Care Facility, Imaging Facility, Laboratory, Post

Acute/Long Term Care Facility: Nursing Home/Skilled Nursing Facility
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team, Facility, Health Plan, Integrated Delivery System, Population: Community, Population: County or City, Population : National, Population : Regional, Population: State
Measure Developer: Ingenix, 950 Winter Street, Suite 3800, Waltham, Massachusetts, 02451

Committee Recommendation for Endorsement: Y-7; N-9; Abstain-0

Conditions/Questions for Developer:

1. Can you give us more information on how repeatability and "consistency" were determined? The results don't appear consistent.
2. Are patients with COPD excluded?
3. How are results reported and interpreted?
4. How would a smaller health plan implement this measure? It seems it might be too complex and burdensome.

Developer Response:

1. Repeatability was demonstrated by programming the measure in SAS code and the Ingenix software and comparing results. Because there are differences in what geographies these health plans are pulling from, variation is expected. But while differences across HCO's are expected, whether the differences are too high or low is difficult to know.
2. Patients are excluded from the asthma episode if they have more costs attributable to COPD than asthma.
3. The main measurement is the O/E ratio metric - the numerator of which is the cost of all the episodes of asthma, and the denominator which is the expected costs.
4. The burden depends on the plan's familiarity with ETGs and similar products, and for those who are just starting out, there is unlimited training involved (i.e. help desk support, etc.). There is another option where Ingenix takes the data and runs it themselves - or uses their PCQ Connect product that prepared the data into report-ready formats.

1.Importance to Measure and Report
1a.High Impact: H-9; M-0; L-0
*TAP Discussion:* The TAP agrees that asthma is a very important health care area to measure.
1b. Resource use/cost problems: H-8; M-1; L-0 ; I-0
*TAP Discussion:* The TAP agrees the Measure demonstrates cost problems and opportunity for improvement.
1c. Purpose clearly described: H-7; M-2; L-0; I-0
*TAP Discussion:* The TAP believes the purpose and objective of the measure are clear.
1d. Resource use service categories consistent and representative: H-7; M-2; L-0; I-0
*TAP Discussion:* The TAP feel this subcriterion has been met.

Overall Importance: Y-16, N-0
Committee Discussion: The Steering Committee agreed that asthma constitutes a high impact healthcare area.

2. Scientific Acceptability of Measure Properties:
2a. Reliability: H-0; M-8; L-1; I-0
2a1.Measure well defined and precisely specified: H-2; M-6; L-1; I-0
*TAP Discussion*: This measure is one that's part of a suite of episodes around diseases and conditions included in Ingenix's episode treatment grouper. This product identifies claims that should be part of an episode of asthma and divides them into year-long segments, looking at asthma as a chronic disease. The episodes are severity adjusted using clinical markers called condition status factors. Anchor episodes, or face-to-face encounters, are merged together into one episode (i.e. "asthma").
2a2. The results are repeatable: H-3; M-5; L-1; I-0
*TAP Discussion:* The TAP didn't understand why Ingenix used three different population samples, rather than taking a portion of the larger population and testing it multiple times. They would like better communication on the approach as well as more detailed depiction of the data. Repeatability was generally determined to be demonstrated adequately, but for the above reasons, some did question the reliability of the measure score.
2b. Overall Validity: H-0; M-6; L-1; I-2
2b1. Evidence is consistent with intent: H-2; M-5; L-1; I-1
*TAP Discussion*: It was unclear to the panel whether Ingenix is actually measuring asthma costs as intended. The determination of what is an asthma cost and what is not isn't transparent. They also agreed that any results are going to be questioned when potentially over 50% of the costs (the pharmacy costs) are not represented. There were suggestions to stratify those health plans that have pharmacy carve-out arrangements.
2b2.Score/Analysis: H-1; M-4; L-2; I-2
*TAP Discussion:* Face validity was determined to be appropriate. The TAP continued to express concern about the exclusion of

pharmacy costs, which were agreed to be a significant component of asthma care. Pharmacy data is not a requirement to get into the episode (for all ETGs).
2b3. Exclusions: H-1; M-7; L-1; I-0
*TAP Discussion:* The TAP was concerned about the lack of transparency regarding which costs were excluded, and why. Confusion existed around what the grouper identified as outliers or exclusions. Winsorizing very high cost episodes, the top 2%, effectively excludes those kinds of patients that would be important to know about. Addition information such as sensitivity analyses would have helped explain the impact of these high cost cases.
2b4. Risk Adjustment: H-1; M-4; L-2; I-2
*TAP Discussion:* The TAP expressed the same concerns regarding the risk-adjustment methodology as they had for previous Ingenix measures. The TAP was apprehensive that because the measure doesn't require use of standardized costs, the playing field is not level and it can't be implemented consistently across organizations if one is using standard and another actual pricing. To examine how refined the risk-adjustment is, R-squares for different severity levels and how they predict resource utilization should be provided.
2b5. Identification of statistically significant/meaningful differences: H-0; M-8; L-0; I-1
*TAP Discussion:* The TAP felt confident in Ingenix's methodology after it was explained.
2b6. Multiple data sources: N/A (using all administrative data)
2c. Stratification for disparities: H-2; M-6; L-0; I-1
*TAP Discussion:* Gender and age can be stratified, but race data is not available.

Overall Reliability: H-1; M-14; L-1; I-0
Overall Validity: H-0; M-8; L-8; I-0
Overall Scientifically Acceptable: Yes [Split vote [Y-8; N-8 (Committee Vote)]
Committee Discussion: The Committee struggled with the circuitous reasoning behind asthma with acute exacerbation being a condition status and then having that condition status factor into the assignment of severity levels. Ingenix defended this methodology by explaining that for all measures, everything related to severity is based on utilization, which, although circular, is the best possible option. The Committee reiterated the TAP's concern that over half of asthma resource use costs are not captured in this measure since pharmacy data is not collected. They expressed unease about the incomparability of entities that have pharmacy data to those that do not.

Usability:
3a. Measure performance results are publicly reported: H-2; M-4; L-2; I-1
*TAP Discussion:* This product is generally used with a suite of ETG's, usually in combination with the pneumonia and COPD measures. There was uncertainty about the measure's usefulness on its own. Since Ingenix can't ascertain if this measure is being used individually the concern from the panel is how the individual measure could be used.
3b. Measure results are meaningful/useful for public reporting and quality improvement: H-0; M-6; L-2; I-1
*TAP Discussion*: The TAP was concerned about the possibility of misinterpretation of results because of the transparency and usability of the results of this measure.
3c. Data and results can be decomposed for transparency and understanding: H-3; M-5; L-1; I-0
*TAP Discussion:* The TAP reiterated their concern of the transparency of the score. Ingenix clarified that there are ways to drill into different aspects of care to see how they might be driving the score.
3d. Harmonized or justification for differences: N/A

Overall Usability: H-0; M-9; L-6; I-1
Committee Discussion: Several Steering Committee members challenged the idea that asthma should be thought of in terms of "episodes," as it is a chronic condition.

4. Feasibility:
4a. Data elements routinely generated during care process: H-7; M-2; L-0; I-0
*TAP Discussion*: The TAP believes this subcriterion has been met; data is a byproduct of care.
4b. Data elements available electronically: H-7; M-2; L-0; I-0
*TAP Discussion*: The TAP agrees this subcriterion has been met; data is available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-1; M-8; L-0; I-0
*TAP Discussion*: The TAP was generally comfortable with the error checks built into the product.
4d. Data collection strategy can be implemented: H-4; M-4; L-0; I-1
*TAP Discussion*: The TAP expressed some concern about the burden this measure would place on a programmer to implement, particularly at smaller health plans.

Overall Feasibility: H-1; M-8; L-7; I-0

Committee Discussion: See Ingenix feasibility discussion above.

900

1608: ETG Based Chronic Obstructive Pulmonary Disease Cost of Care Measure (COPD) (Ingenix)

Description: The measure focuses on resources used to deliver episodes of care for patients with COPD.
COPD episodes are defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating COPD. A number of resource use measures are defined for COPD episodes, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per episode and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons.
As requested by NQF, the focus of this submission is for COPD episodes and will cover both measures at the COPD base and severity level and also a COPD composite measure where COPD episode results are combined across COPD severity levels. At the most detailed level, the measure is defined as the base condition of COPD and an assigned level of severity (e.g., resources per episode for COPD, severity level 1 episodes). Composite measures can then be created using these measurement units to meet a specific need. For example, a composite measure for COPD is derived by combining COPD episode results across COPD severity levels. Appropriate risk adjustment is applied to support comparisons (e.g., for physician measurement, adjusting for a physician's mix of COPD episodes by severity level when supporting a COPD composite comparison). The focus of this measure is on COPD. However, COPD episode results could also be included in a "pulmonary" "chronic care", or other clinical composite for a physician, combining episodes in clinical areas similar to COPD. Further, an "overall" composite for a physician can be created, again by aggregating episode results across appropriate conditions and severity levels and applying proper risk adjustment when making comparisons.
Resource Use Type: Per episode
Data Type: Administrative claims, Other
Resource Use Service Categories:
Inpatient services: Inpatient facility 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 : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinic/Urgent Care, Ambulatory Care : Clinician Office, Emergency Medical Services/Ambulance, Home Health, Hospice, Hospital/Acute Care Facility, Imaging Facility, Laboratory, Post Acute/Long Term Care Facility: Nursing Home/Skilled Nursing Facility
Level of Analysis: Clinician : Group/Practice, Clinician : Individual, Clinician : Team, Facility, Health Plan, Integrated Delivery System, Population: Community, Population: County or City, Population : National, Population : Regional, Population: State
Measure Developer: Ingenix, 950 Winter Street, Suite 3800, Waltham, Massachusetts, 02451

Committee Recommendation for Endorsement: This measure did not pass the scientific acceptability criterion, and is not recommended for endorsement.

Conditions/Questions for Developer:
1. What was the clinical logic of using 180 days, particularly since your Asthma measure had used 365 days, and both are similar chronic conditions?

Developer Response:
1. We will have to examine that further.

1. Importance to Measure and Report

1a.High Impact: H-7; M-0; L-0; 1-0
*TAP Discussion:* The TAP agreed Ingenix did well with articulating the high impact of COPD.
1b. Resource use/cost problems: H-7; M-0; L-0; I-0
*TAP Discussion:* The TAP believe that COPD represents a resource use issue that can be addressed.
1c. Purpose clearly described: H-7; M-0; L-0; I- 0
*TAP Discussion:* The TAP feel the purpose and objective are clear.
1d. Resource use service categories consistent and representative: H-6; M-1; L-0; I-0
*TAP Discussion:* The TAP believes this subcriterion has been met.

Overall Importance: Y-16, N-0
Committee Discussion: There was unanimous agreement that asthma constitutes a high impact area of healthcare.

2. Scientific Acceptability of Measure Properties:

2a. Overall Reliability: H-4; M-3; L-0; I-0
2a1.Measure well defined and precisely specified: H-4; M-3; L-0; I-0
*TAP Discussion*: The TAP discussion focused around the clinical logic around the timeframes chosen.
2a2. The results are repeatable: H-5; M-2; L-0; I-0
*TAP Discussion:* The TAP agrees that reliability for this measure is similar to the previously discussed Ingenix asthma measure.
2b. Overall Validity: H-0; M-7; L-0; I-0
2b1. Evidence is consistent with intent: H-2; M-5; L-0; I-0
*TAP Discussion*: The TAP believes this subcriterion has been met.
2b2.Score/Analysis: H-0; M-7; L-0; I-0
*TAP Discussion:* The TAP remained concerned about Ingenix's testing method for customization, the inability to compare actual versus standardized prices, and the high level of pharmacy exclusions.
2b3. Exclusions: H-1; M-6; L-0; I-0
*TAP Discussion:* There are no clinical exclusions, only administrative ones. The TAP felt it was unclear how tie-breaking logic works and noted that it was not specified in the submission how COPD and asthma ETG's interact.
2b4. Risk Adjustment: H-0; M-4; L-3; I-0
*TAP Discussion:* While Ingenix had a nice description of how they developed their risk-adjustment approach, the panel would have liked to see more description of the modeling presented in the submission.
2b5. Identification of statistically significant/meaningful differences: H-0; M-7; L-0; I-0
*TAP Discussion:* The TAP questioned whether the practical significance of the measure since it is a relative cost ratio.
2b6. Multiple data sources: N/A (using all administrative data)
2c. Stratification for disparities: H-2; M-5; L-0; I-0
*TAP Discussion:* Only gender and age are stratified for. Race data is not available.

Overall Reliability: H-3; M-10; L-2; I-0
Overall Validity: H-1; M-5; L-9; I-0
Overall Scientifically Acceptable: Yes [Y-5; N-10 (Committee Vote)]
Committee Discussion: The Steering Committee appreciated the change Ingenix made to the measure's timeframe at the TAP's suggestion, from 180 to 365 days, to remain consistent with the asthma measure. It was felt the analysis of scientific acceptability for this measure would generally reflect the same analysis for measure 1560 Asthma.

Usability:
3a. Measure performance results are publicly reported: H-0; M-7; L-0; I-0
*TAP Discussion:* The TAP expressed doubts regarding whether the measure could be implemented in a user-friendly manner.
3b. Measure results are meaningful/useful for public reporting and quality improvement: H-0; M-7; L-0; I-0
*TAP Discussion:* The panel agreed that measure provides useful information for individual health plans. However, they expressed concern about how useful it would be to compare across health plans, due to the fact that standardized pricing is not required.
3c. Data and results can be decomposed for transparency and understanding: H-3; M-4; L-0; I-0
*TAP Discussion:* It was agreed that previous discussions regarding Ingenix transparency would also apply to this measure.
3d. Harmonized or justification for differences: N/A

Overall Usability: This measure did not pass the scientific acceptability criterion. As a result, the Committee did not discuss usability.

4. Feasibility:
4a. Data elements routinely generated during care process: H-5; M-2; L-0; I-0
*TAP Discussion*: The TAP believes this subcriterion has been met; data is a byproduct of care.
4b. Data elements available electronically: H-7; M-0; L-0; I-0
*TAP Discussion*: The TAP believes this subcriterion has been met; data available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-3; M-4; L-0; I-0
*TAP Discussion*: The TAP is comfortable that Ingenix can accurately identify inaccuracies and errors.
4d. Data collection strategy can be implemented: H-6; M-1; L-0; I-0
*TAP Discussion*: The TAP believes this subcriterion has been met.

Overall Feasibility: This measure did not pass the scientific acceptability criterion. As a result, the Committee did not vote on feasibility.

## Candidate Consensus Standards with No Committee Consensus

The Committee was unable to come to consensus on one candidate consensus standard.

The following evaluation summary table summarizes the results of the TAP's and Committee's evaluation of and voting on the candidate consensus standard that did not draw Committee consensus. A hyperlink is provided in summary table to the detailed measure specifications. To access the meeting transcripts and recordings in which this measure is discussed, refer to the project web page.



## Evaluation Summary—Candidate Consensus Standard with No Committee Consensus

| 1595: ETG Based Diabetes Cost of Care Measure (Ingenix) |
|---|
| Description: The measure focuses on resources used to deliver episodes of care for patients with Diabetes. Diabetes episodes are defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating diabetes. A number of resource use measures are defined for diabetes episodes, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per episode and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons. The focus of this submission is for Diabetes episodes and will cover both measures at the Diabetes base and severity level and also a Diabetes composite measure where Diabetes episode results are combined across Diabetes severity levels. At the most detailed level, the measure is defined as the base condition of diabetes and an assigned level of severity (e.g., resources per episode for diabetes, severity level 1 episodes).<br>Resource Use Measure Type: Per episode<br>Data Source: Administrative claims, Other<br>Resource Use Service Category: Inpatient services: Inpatient facility 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<br>Care Setting: Ambulatory Care: Ambulatory Surgery Center (ASC), Ambulatory Care: Clinic/Urgent Care, Ambulatory Care: Clinician Office, Home Health, Hospice, Hospital/Acute Care Facility, Imaging Facility, Laboratory<br>Level of Analysis: Clinician: Group/Practice, Clinician: Individual, Clinician: Team, Facility, Health Plan, Integrated Delivery System, Population: Community, Population: County or City, Population: National, Population : Regional<br>Measure Developer: Ingenix |
| Committee Recommendation for Endorsement: Y-7; N-7; Abstain -0 (re-vote) [Y-11; N-7; Abstain-0 (initial vote)] |
| 1. Importance to Measure and Report:<br>1a.High Impact: H-9 ; M-0 ; L-0 ; I-0<br>*TAP Discussion:* The TAP believes this is a high cost, impact aspect of healthcare; this subcriterion has been met.<br>1b. Resource use/cost problems: H- 3 ; M-6 ; L-0 ; I-0<br>*TAP Discussion:* The TAP would have liked to see more evidence of provider variation and other types of variation in treating diabetes in addition to the regional variation.<br>1c. Purpose clearly described: H- 4 ; M-5 ; L-0 ; I-0<br>*TAP Discussion:* The TAP believes that the intent provided not specific to this diabetes measure, it is a very general statement.<br>1d. Resource use service categories consistent and representative: H- 9 ; M-0 ; L-0 ; I-0 |

1595: ETG Based Diabetes Cost of Care Measure (Ingenix)

*TAP Discussion:* The TAP believes this subcriterion has been met.

Overall Importance: Y-18, N-0
Committee Discussion: The Steering Committee believes this is a high impact area that should be measured; this subcriterion has been met.

2. Scientific Acceptability of Measure Properties:
2a1. Well defined/precise specifications: H- 5 ; M-3 ; L-1 ; I-0
*TAP Discussion:* Specifications for co-morbidities, severity levels, etc. are not clear. It is unclear if severity ratings are weighted based on services of comparable cost. Only costs that are mapped back to the diabetes code are counted in the episode. The measure is stratified by severity level not clinical condition. Concerns about how patients with pharmacy benefit (or who run out of pharmacy benefit) are compared to those with full pharmacy benefit.
2a2. Reliability testing: H- 7 ; M-1; L-0 ; I-0
*TAP Discussion:* Demonstration of internal consistency was presented to demonstrate reliability. The Committee requested additional reliability tests in during maintenance. Additional detail in terms of the r2 of the risk adjustment model and calibration results was requested.
2b1. Specifications consistent with resource use/cost problem: H- 1 ; M-6 ; L-1 ; I-0
*TAP Discussion:* TAP was unclear on whether diabetes education codes were included in the specifications?
2b2. Validity testing: H- 4 ; M-3; L-0 ; I-1
*TAP Discussion:* The TAP believes adequate validity testing information provided. More robust methods should be considered in future evaluations.
2b3. Exclusions: H-0; M-7 ; L1 ; I-0
*TAP Discussion:* TAP was unclear on how exclusions were identified.
2b4. Risk adjustment : H-0 ; M-4 ; L-4 ; I-0
*TAP Discussion:* The TAP was concerned about the inability to distinguish between complications and comorbidities.
2b5. Identification of statistically significant/meaningful differences: H- 0 ; M-4 ; L-4 ; I-0
*TAP Discussion:* Insufficient evidence that the sample size threshold and analysis at the physician level is meaningful at that level. Unclear how the 30 sample size was selected.
2b6. Multiple data sources: N/A
2c. Stratification for disparities: H- 0 ; M-0 ; L-0 ; I-0; N/A-9
*TAP Discussion:* Due to the limitations in the administrative claims data, at this time the measure does not stratify for disparities.

Overall Scientifically Acceptable: Yes [Y-10; N-8 (Committee Vote)]
Committee Discussion: As an introduction to the measure, the developer summarized their responses to the TAP concerns including that the diabetes education codes have been confirmed and are included in the specifications. Similar to the TAP, the Committee expressed concern about the minimum sample size guideline suggesting 30 cases per physician; the Committee questioned how this number was identified and if any statistical analysis was performed to support this guideline. In response to this concern, the developer explained that this sample size was borrowed from previous work done by NCQA on resource utilization and stated that from their perspective, while sample size can be important, ensuring results are statistically significant is more important. The Committee also requested explanation of the attribution model, finding that it was very complex, and questioned of the total sample from their analysis, what percent of physicians have a minimum sample size of 30. The developer explained that the attribution model seeks to identify the highest number of contacts between the physician and the patient related to diabetes; in case of a tie, the provider with the highest actual cost gets attributed the episode. Another concern identified by the Committee relates to how the measure captures costs related to the sequela of diabetes (e.g., renal disease, eye disease, CHF); the measure as presented does not currently account for these costs as they trigger alternate episodes. There was also discussion on how this measure (or measures like it) might be paired with quality (process) measures, as it measures resource use and adjusts for conditions *before* care is provided. The Committee also spent some time discussing and trying to understand the episode trigger mechanisms, such as when a patient enters the episode in the middle of the 12-episode; in this case the episode is marked incomplete. There was a question to the developer about what percentage of the claims was higher or lower than expected. The developer was unable to answer the question off hand but will get back to the Committee with this information. The issue of mental health and pharmacy carve outs was a prevalent issue throughout the discussion of these measures. For this measure mental health is not stratified for when it is carved out.

3. Usability:
3a. Measure performance results are publicly reported: H- 0; M-1 ; L-1; I-6
*TAP Discussion:* The usability information submitted is not specific to diabetes, but for all Ingenix measures. TAP expressed concerns

| 1595: ETG Based Diabetes Cost of Care Measure (Ingenix) |
|---|
| with the availability of this data to the public and requested clarification from NQF on what is required for "public reporting". The NQF CSAC and BOD continue to discuss this issue; NQF staff will continue to filter any new information on the refining of this policy to the TAP to facilitate final ratings of this usability criterion.<br>3b. Measure results are meaningful/useful for public reporting and quality improvement: H- 0 ; M-4 ; L-2 ; I-2<br>*TAP Discussion:* The usability information submitted is not specific to diabetes, but for all Ingenix measures.<br>3c. Data and results can be decomposed for transparency and understanding: H- 1 ; M-2 ; L-5 ; I-0<br>*TAP Discussion:* The usability information submitted is not specific to diabetes, but for all Ingenix measures. Challenges for the use of this measure include, complexity, lack of specificity in specifications. The TAP agrees it is difficult to assess the extent of which the measure can be decomposed as currently specified.<br>3d. Harmonized or justification for differences: H-0 ; M-0 ; L-0 ; I-0; N-9<br>*TAP Discussion:* The usability information submitted is not specific to diabetes, but for all Ingenix measures. |
| Overall Usability: H-0; M-9; L-6; I-3<br>Committee Discussion: While there is a transparency website for physicians to go to in order determine what a score means, it may take a lot of time to do this. The Steering Committee questioned whether this is a reasonable expectation and adequately demonstrates transparency. Other concerns raised by the Steering Committee were related to the attribution model and how the complexity of the methodology might impact how understandable the measure construction and results are. Because this measure is part of an episode grouper and is not used in isolation as an individual measure, the information the developer was able to present on its current use is not specific to the diabetes episode, but the product as a whole. |
| 4. Feasibility:<br>4a. Data elements routinely generated during care process: H- 8 ; M-0 ; L-0 ; I-0<br>*TAP Discussion:* The TAP agrees this subcriterion has been met; measures rely on administrative data.<br>4b. Data elements available electronically: H-8 ; M-0 ; L-0; I-0<br>*TAP Discussion:* The TAP agrees this subcriterion has been met; administrative data are in electronic format.<br>4c. Susceptibility to inaccuracies/ unintended consequences identified: H-2 ; M-2 ; L-4 ; I-0<br>*TAP Discussion:* The TAP does not feel this subcriterion was adequately met; there are current issues identified with specifications could result in inaccuracies and errors.<br>4d. Data collection strategy can be implemented: H- 5 ; M-2 ; L-1 ; I-0<br>*TAP Discussion:* The TAP agrees that barriers to use are minimal. (NQF Note: This is prior to the submission of product pricing information reviewed only by the Steering Committee). |
| 4. Feasibility: H-2; M-8; L-8; I-0<br>*Committee Discussion:* See Ingenix feasibility discussion above. |

## WITHDRAWN BY DEVELOPER

The 12 measures listed below were withdrawn from the Cycle 2 review process by the developers for further refinement and testing.

### Pulmonary

- (1577) Episode of care for patients with asthma over a one year period (ABMS-REF)
- (1581) Episode of care for patients with stable chronic obstructive pulmonary disease over a one year period (ABMS-REF)
- (1582) Episode of care for patients with unstable chronic obstructive pulmonary disease over a one year period (ABMS-REF)
- (1587) Episode of care for ambulatory pneumonia (ABMS-REF)

- (1588) Episode of care for community acquired pneumonia hospitalization (ABMS-REF)

### Cancer

- (1578) Episode of care for 60-day period preceding breast biopsy (ABMS-REF)
- (1579) Episode of care for cases of newly diagnosed breast cancer over a 15 month period (ABMS-REF)
- (1583) Episode of care for 21-day period around a colonoscopy (ABMS-REF)
- (1584) Episode of care for treatment of localized colon cancer (ABMS-REF)

### Bone/Joint

- (1585) Episode of care for simple, non-specific lower back pain (acute and subacute) (ABMS-REF)
- (1586) Episode of care for acute/subacute lumbar radiculopathy with or without lower back pain (ABMS-REF)
- (1610) ETG based low back pain resource use measure (Ingenix)

## ADDITIONAL CONSIDERATIONS

As the first NQF resource use measure review and evaluation process concludes, there is a great opportunity to reflect on and provide recommendations for future efforts in this area. While resource use measurement has been used in the commercial sector for many years, the emerging interest in using these measures for public reporting and payment initiatives further highlights the need for efforts such as this to explore the complexities and potential challenges for multiple applications. The Committee was asked to provide guidance to the field on how measure of resource use can be improved and provide special considerations for developing measures for the Medicare program. Additionally, members provided insight on how measures of resource use may be linked with quality measures to provide a true assessment of efficiency. Through this exercise, the Committee offered recommendations in several areas. In doing so, several new principles emerged.

### 1) Submitting and Evaluating Resource Use Measures

*Emerging Principle 1: While guidelines in measure components may be acceptable for internal quality improvement, to promote measurement for comparison across entities nationally, the entire resource use measure construct should be standardized in the form of specifications.*

*Emerging Principle 2: The risk adjustment model applied to the measure should be specific to the intended population.*

*Emerging Principle 3: The factors in the risk adjustment model and severity model should be confirmed to be a contributor to the outcome of the measure.*

*Emerging Principle 4: In addition to statistical significance, justification of the variables used in the risk-adjustment model should be provided based on either clinical relevance or evidence in the literature.*

### 2) Reliability and Validity Testing

*Emerging Principle 5: To demonstrate reliability of a resource use measure, developers can focus on precision of the measure score.*

*Emerging Principle 6: When there is such limited variability in a data set that it does not adequately distinguish performance differences among providers, reliability cannot simply rely on confidence intervals; sample size should also be included in the reliability assessment.*

*Emerging Principle 7: The gold standard approach to determining the validity of data elements based on administrative claims data in resource use measures is to assess the agreement of claims data with source of the data elements in the chart.*

### 3) Data Quality and Comprehensiveness

*Emerging Principle 8: Data sets used to measure resources should be as comprehensive as possible. Efforts to obtain clinical and carved-out data (e.g., pharmacy, behavioral health) should be made to ensure the data set used to calculate resource use is robust, complete, and representative.*

*Emerging Principle 9: Measure scores calculated and reported using data with carve-outs should be labeled as such.*

*Emerging Principle 10: Comparisons of entities with and without carved-out data is inappropriate.*

*Emerging Principle 11: If a measure is intending to measure a clinical condition that encompasses a predominant portion of its costs in pharmacy claims, consider whether costs should be measured at all in the absence of these data. It is the developers' responsibility to conduct an analysis to determine whether the lack of these data invalidates the measure score or comparisons.*

**4) Measuring Cost and Resource Use in the Medicare Population**

*Emerging Principle 12: A patient-centered approach should be used to describe the interaction of conditions (and episodes) in the development of resource use measures for the Medicare population.*

**5) Linking Quality and Cost to Develop Measures of Efficiency and Value**

*Emerging Principle 13: Efficiency measurement approaches should be patient-centered, building upon previous efforts such as the NQF Patient-Centered Episodes of Care (EOC) Efficiency Framework.*

Recommendations and emerging principles are discussed below:

**1) Submitting and Evaluating Resource Use Measures**

In an effort to minimize the confusion in the submission and evaluation processes, the Committee identified areas within the resource use measure specification modules that should be clarified so that the developer understands the information that is required for the measure to be considered fully. The Committee recognized that in an effort to improve the clarity of the measure submissions, there should also be attention paid to how new submission requirements will affect the burden on the developer to submit measures for consideration. While there are some areas of the submission that will need additional information and more clarity, there may be other pieces of information that may not be required.

*Emerging Principle 1: While guidelines in measure components may be acceptable for internal quality improvement, to promote measurement for comparison across entities nationally, the entire resource use measure construct should be standardized in the form of specifications.*

The data protocol module components were framed as flexible user instructions for missing data, data inclusion and exclusions, and data cleaning that could be submitted as specifications or

guidelines. All of the measure submissions and all of the data protocol components were submitted as guidelines. Allowing for flexibility in this module led to some discomfort for the experts specifically related to handling missing data. Ensuring that the data used to run the resource measures are complete and representative is a critical first step to generating valid measure results. Allowing flexibility in these steps could allow for errors and inconsistent implementation of the data cleaning and data preparation steps. As such, the Committee recommends that the steps within the data protocol module be submitted as specifications going forward. Specifically, it should be indicated explicitly in the submission whether it is acceptable to implement the measure using a data set with carve-outs. Data cleaning steps should be explicitly stated as specifications.

Likewise, in the reporting module, while the attribution approach could be submitted as specifications or guidelines, the Committee was very concerned with how the models reviewed might be applied, even as guidelines. While there was not widespread agreement on any of the attribution approaches reviewed, the Committee recognized that there must be some attribution approach employed with the use of these measures to facilitate actionable measurement since many states and healthcare systems may require varied approaches for their unique market. This highlights the need for more discussion on how, if at all, attribution approaches should be evaluated in this process where the goal is to endorse standardized approaches to measurement.

*Emerging Principle 2: The risk-adjustment model applied to the measure should be specific to the intended population.*

After reviewing various risk-adjustment approaches presented in the measures submitted to this project, the Committee agreed measure developers need to demonstrate that the specified risk models are appropriate for the target population. For instance, if the hierarchical condition category (HCC) model is used to measure a commercial population, developers need to demonstrate that it is appropriate for use outside of a Medicare population. The Committee agreed that risk models have unique weights for comorbidities and may not include all relevant conditions (for example, pregnancy) when the risk-adjustment model is used outside of the population in which it is calibrated. Measure developers have the burden of demonstrating appropriateness through R-

squared values and through a detailed clinical and statistical explanation of how variables were added to the risk model. Additional research is needed in this area to explore how various risk-adjustment approaches change the relative ranking of providers in terms of resource use and how the use of clinically enhanced administrative data may impact measure scores and the selection of factors added to the risk-adjustment models.

*Emerging Principle 3**: The factors in the risk-adjustment model and severity model should be confirmed to be a contributor to the outcome of the measure.*

*Emerging Principle 4: In addition to statistical significance, justification of the variables used in the risk-adjustment model should be provided based on either clinical relevance or evidence in the literature.*

The Committee agreed that measure developers need to demonstrate that variables included in the risk-adjustment model are not simply selected based on their statistical explanatory power, but rather, risk factors are well documented in clinical evidence. When variables are chosen for inclusion in the risk-adjustment model, developers are responsible for demonstrating a relationship to the outcome of the measure (i.e., resource use). Additional detail through a sensitivity analysis including various risk-adjustment variables can be provided in future evaluations to demonstrate the effect of variables included in the final risk-adjustment approach.

### 2) Reliability and Validity Testing

The cumulative experience of the multiple TAPs and the Resource Use Steering Committee demonstrated that resource use measures developers are at various levels of measure testing sophistication.

Measures submitted as resource use national consensus standards need to improve their level of sophistication of reliability and validity testing. To balance the developer burden of testing for the initial evaluation of resource use measures with providing the experts the information needed to make a valid conclusion about reliability and validity, the TAPs and Steering Committee agreed that the scope of testing may be on a relatively small scale for initial endorsement. The

Committee agreed further analysis by all developers would be required to support continued endorsement at the time of review in order to maintain NQF endorsement.

Reliability and validity testing is included in the NQF evaluation criteria, and NQF allows flexibility in the specific methods used in testing to allow measure developer flexibility. The Committee evaluated: 1) the scope of testing, 2) what tests of reliability and validity could be performed, and 3) how to weigh the results of this testing. The Steering Committee interpreted testing results within the unique context of the specific measure under review.

***Reliability testing***

The NQF evaluation criteria states that reliability testing should demonstrate that the data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period, or that the measure score is precise. The Committee agreed that developers can demonstrate that the measure score is precise by demonstrating an adequate ratio of signal to noise, or how well one can confidently distinguish the performance of one physician from another.[12] The signal is ability of the measure to identify real differences in performance, whereas the noise attributed to measurement error. Demonstrating reliability in this context relies on three major drivers: sample size, differences among physicians, and random variation in the measure scores, or measurement error.[13]

*Emerging Principle 5: To demonstrate reliability of a resource use measure, developers can focus on precision of the measure score.*

Reliability at the data element level of resource use measures submitted to this project relied on administrative claims and by virtue of their design as coded programs were repeatable. However, the Committee clarified that while coded programs may be repeatable at the data element level, measure developers need to demonstrate adequate validity testing at the data element level.

*Emerging Principle 6: When there is such limited variability in a data set that it does not adequately distinguish performance differences among providers, reliability cannot simply rely on confidence intervals; sample size should also be included in the reliability assessment.*

Reliability of resource use measures at the measure score level needs to demonstrate that the measure score is precise. Providing confidence intervals in measure reporting does not sufficiently demonstrate reliability of the measure.

NQF does not prescribe what tests of reliability could be performed, specific thresholds for results, or how to weigh the results of this testing since an evaluation should account for the context of the test, measure, and the data source. The evaluation should incorporate both empirical evidence and expert judgment to evaluate whether the specific measure under evaluation by the Committee has sufficiently demonstrated reliability through the measure submission.

***Validity testing***

*Emerging Principle 7: The gold standard approach to determining the validity of data elements based on administrative claims data in resource use measures is to assess the agreement of claims data with source of the data elements in the chart.*

Since the entire dataset may not be available for such validation, applying the resource use measure to a simulated data set that should return known values of the data elements and scores may be used. With either approach, when the results obtained for the resource use measure do not match known values in the simulated data set or the abstracted data, an analysis should be conducted to determine the source of error.[15] If the error is related to the measure specifications, including code lists, clinical or construction logic, and computer readable programming language, the measure specification should be corrected before submitting for endorsement.

The NQF criteria state that validity testing must demonstrate that the measure data elements are correct or that the measure score correctly reflects the cost of care or resources provided, adequately distinguishing high and low resource use. Developers must demonstrate measures have undergone sufficient validity testing demonstrating that the resource use measure actually measures what it claims to measure. If only face validity is addressed, it must be assessed systematically. The Committee recommended that validity testing be demonstrated 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. Correct conclusions about resource use can be made when validity tests demonstrate that claims used in the measure accurately reflect information in the charts of a representative sample of patients. The Committee considered that most developers submitting to this project may not have direct access to chart abstracted data; however, additional efforts are strongly recommended to ensure data elements used to develop the resource use measures are valid.

### 3) Data Quality and Comprehensiveness

In an effort to address some of the underlying global issues affecting the use of administrative claims data for the purposes of measuring resource use, the Committee identified several areas in which healthcare stakeholders might engage and support additional efforts to improve the ability of resource use measures to capture all resource use fully. In doing so, a few new principles for resource use measurement emerged.

*Emerging Principle 8: Data sets used to measure resources should be as comprehensive as possible. Efforts to obtained clinical and carved-out data (e.g., pharmacy, behavioral health) should be made to ensure the data set used to calculate resource use is robust, complete, and representative.*

*Emerging Principle 9: Measure scores calculated and reported using data with carve-outs should be labeled as such.*

*Emerging Principle 10: Comparisons of entities with and without carved-out data is inappropriate.*

A major concern of the Committee throughout the evaluation process was the impact of carve-out arrangements on accurately capturing resources used. While there are some systems that are able to recapture these data from the outsourced entities, others do have this capability. Furthermore, measure results derived for entities with carve-out arrangements should be labeled as such to prevent comparison between entities with and without such carve-out arrangements. The measures received during this project were specified using administrative claims data only. The use of administrative claims data presents certain limitations for measuring resource use performance, limitations that are present in quality performance measurement as well. Primarily

the reliance of resource use measures on administrative claims data to count resources, or dollars spent, captures only the output on behalf of the provider—not the costs to the patient, nor the costs or resources for which there are no administrative codes. Recognizing this as a limitation of the data available to measure these types of resources, the Committee recommended that future efforts in resource use measurement focus not only on the costs to the provider, but to the user as well, through identifying those resources that are important to measure and determining how to capture this data. The Committee recognized that while administrative data are the primary data source used for measuring resources at this time, there is opportunity to integrate the data gathered through EHRs and other clinical data to measure resource use.

Since resource use measurement is a priority, efforts should be made to ensure the necessary data are available for accurate measurement. However, there are significant challenges to determining where the responsibility lies to ensure data are complete and the ways in which important but sensitive information is shared. For a number of measures submitted for evaluation in this project, the instructions within the data protocol module suggest that the measure implementer is responsible for ensuring data are complete and representative. The Committee agreed however that measure implementers often do not have the resources or technical expertise to audit data before use. Future efforts should explore a potential role for large data aggregators to identify thresholds and set standards for data quality.

*Emerging Principle 11: If a measure is intending to measure a clinical condition that has a predominant portion of its costs in pharmacy claims, consider whether costs should be measured at all in the absence of these data. The developer is responsible for determining whether the lack of these data invalidates the measure score or comparisons.*

When developing resource use measures, careful consideration should be given to whether the importance of measuring resources/costs in an area outweigh the limitations of the data. For some conditions, the lack of robust data could distort the measure output. For example, to measure the resources for asthma patients where greater than 40 percent of the resource use is pharmacy related, data sets without pharmacy data are inherently misleading in providing useful

insight into the cost of asthma care.[16] For acute or procedural episodes (e.g., hip replacement) where the care is more standardized (e.g., pre- and post-surgical antibiotics) pharmacy and mental health data do not account for a major portion of the resource use and thus administrative data, and carve-out issues may not have a tremendous impact on the measure results.

### 4) Measuring Cost and Resource Use in the Medicare Population

Measures evaluated in this project were mainly specified for the commercial population; however, the Steering Committee identified areas of consideration for organizations developing resource use measure for Medicare beneficiaries. Resource use measures for the Medicare population will have to consider multiple co-occurring conditions, as well as multiple sites where beneficiaries seek care and resource use at the end of life. This guidance was provided with the understanding that there will be an urgent need for measures specified for the Medicare population for use in bundled payment demonstrations, physician feedback reporting programs, and value-based purchasing programs.

An important consideration in resource use approaches developed for the Medicare population is the presence of multiple co-occurring conditions. The Committee considered that more than half of all beneficiaries were treated for five or more conditions, accounting for three-fourths of total Medicare spending.[17] In 2002, more than 92.2 percent of all Medicare healthcare spending was incurred by beneficiaries with three or more conditions during the measurement year.[18] In the Committee's discussion of the approach to measure resource use in patients with multiple co-occurring conditions, they concluded that cost estimates should be based on the time and attention a provider should be reasonably expected to deliver on a patient's multiple co-occurring conditions beyond the acute disease and its immediate complications for which the patient sought care.

*Emerging Principle 12: A patient-centered approach should be used to describe the interaction of conditions (and episodes) in the development of resource use measures for the Medicare population.*

Episode approaches attempting to assign claims to specific episodes should create a transparent hierarchy with rules to assess resource use in the Medicare population accurately. One approach the Committee suggested would allow flexibility in the assignment of individual claims to a single episode or to multiple open episodes. This patient-centered approach could allow an individual office visit for evaluation and management to be assigned to multiple episodes.

Efforts to develop resource use measures for the Medicare population should consider the NQF consensus measure framework for assessing the efficiency of care for individuals with multiple chronic conditions (MCCs). MCC framework guiding principles include promoting shared accountability with members of the healthcare system, a multi-dimensional measure approach that incorporates various types of measures, a focus on shared decision making in concordance with a patient's preferences, and prioritization of measures across time that are most relevant to achieving desired outcomes as determined by the care plan.

The Committee recognized the cost contribution of individual conditions to the total cost of managing a beneficiary may vary differently depending on other conditions present in each beneficiary. The following classification system of four types of overlapping episodes helps to illustrate the patterns of treatment discussed by the Committee: (1) linear additive episodes, (2) interactive episodes-cost increasing, (3) interactive episodes-cost savings, and (4) dominant episodes.[19] Linear additive episodes occur when the patterns of illness are not overlapping, and episodes can be considered independent of one another. For example, a fracture of the radium and strep throat would be considered independent of one another.[20] Interactive episodes can be cost increasing when there are two or more conditions in which the presence of multiple conditions increases the level of resources required to treat all of the conditions. An example would include the treatment of diabetes in the presence of obesity.[21] Under this condition, the cost of the combined condition is more than the sum of the individual parts. Interactive episodes can also be cost saving since the cost of treating overlapping conditions is not likely to require significantly different resources (e.g., the treatment of otitis media and bronchitis).[22] Finally, dominant and mild disease combinations in which the presence of a dominant disease episode becomes the principle focus of care (e.g., the treatment of end-stage renal disease in the presence

of mild asthma). These methods for overlapping episodes should be considered in developing approaches for assessing resource use in the Medicare population.

The nature of the interaction between chronic and acute conditions should be considered when developing resource use measures. When developing measures to assess the resource use for a chronic condition, the resource use for an acute complication for that condition should be considered. The Committee considered the example of misinterpreting lower CAD resource use as better performance when, in fact, a per-capita assessment may demonstrate higher resource use. The higher resource use may be derived from higher rates of AMI in the measured population due to poor CAD management.

Additional efforts are needed to propose alterative attribution approaches to encourage team-based care along the patient episode of care. Resource use measures developed for the Medicare population should also consider that beneficiaries often seek care from multiple sites. The typical Medicare beneficiary sees two primary care physicians and five specialists working in four different practices.[23] The Committee discussed how current attribution models assign treatment of the patient to an individual provider based on the number of visits or the highest proportion of costs. However, in a patient-centered model all providers who treat the patient should have responsibility for the care delivered.

Episode-based approaches for the Medicare population should carefully consider their approach to dealing with end of life (EOL). Simply including EOL patients in estimates of episode-based resource use has the potential to introduce inappropriate incentives. Resource use measures that include EOL patients should be reported with balancing mortality measures to ensure that providers are not inadvertently reported as providing more efficient care when they have higher rates of mortality. On the other hand, with resource use during the last year of life accounting for more than a quarter of Medicare payments,[24] EOL patients should not be excluded from the analysis of resource use. Future evaluation of resource measures for the Medicare population should consider how measure developers handle EOL patients in profiling providers.

### 5) Linking Quality and Cost to Develop Measures of Efficiency and Value

Developing measures of efficiency and value is critical to reducing the healthcare cost growth rate. In a first step toward developing efficiency measures, resource use measures must demonstrate they are important to measure, have scientifically acceptable properties, and are usable and feasible. Resource use measures that meet these criteria may be used in conjunction with quality measures to assess efficiency. The Steering Committee reflected on the mechanism and future work needed to achieve this goal.

*Emerging Principle 13: Efficiency measurement approaches should be patient-centered, building upon previous efforts such as the NQF Patient-Centered Episodes of Care (EOC) Efficiency Framework.*

Measures components may need to be aligned between quality and resource use measures. Components that may be aligned include the handling of exclusions, level of analysis, risk adjustment, and stratification approach. For example, the Committee recommended that quality and resource use measures be aligned in terms of their inclusion and exclusion criteria to ensure similar populations are being measured in both the resource use and quality performance.

The Committee recommended future work to define the type of quality and resource use measures that can be used to assess quality. Considerations should include the measure type (e.g., outcome, process, patient experience), measurement period (e.g., single point in time, spanning the measurement year), and the number of quality measures that should be paired with a resource use measure. The Committee also considered that quality measures may be used to monitor for underuse on needed care. Assessments of efficiency will require careful consideration of the mechanism in which quality and resource use measures are linked.

Future efforts should explore approaches to ensure that providers are benchmarked on cost performance against providers with similar or better quality performance. Benchmarking cohorts of providers based on quality performance allows for accurate interpretation of cost. Specifically this method ensures that the resource use performance is compared to only those providers with equal or higher quality performance.[25,26] When available, the Committee agreed that outcome and patient experience of care measures with sufficient reliability (signal to noise) and validity should be selected to assess efficiency.[27]

## NEXT STEPS

This project enabled first-hand experience in reviewing and understanding some of the various approaches for measuring resources and costs in healthcare, and while many lessons were learned, there is still abundant opportunity to apply the principles and recommendations that emerged from this work in future efforts. Ongoing work in the public sector to develop a public episode grouper for the Medicare population and explore ways to measure efficiency using a patient-centered approach will be the focus of future NQF efforts in this area. Additionally, using the recommendations from the Committee on improving the evaluation process, updates to the NQF resource use measure submission forms and evaluation criteria will be explored as we continue to enhance the endorsement process for measure submitters and evaluators.

## 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, *JAMA*, 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 RF, 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 Eselius LL, Mental health carve-outs: effects and implications, *Medical Care Research and Review*, 1999;56(suppl 2):37.

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

12. Scholle, S. H., J. Roski, et al. (2008). "Benchmarking physician performance: reliability of individual and composite measures." The American Journal of Managed Care 14(12): 833.

13. Adams J, McGlynn E, et al., *Incorporating statistical uncertainty in the use of physician cost profiles*, *BMC Health Services Research*, 2010;10(1): 57.

14. Ibid.

15. NQF Testing Task Force Report (2011).

16. Bahadori K, Doyle-Waters, et al., Economic burden of asthma: a systematic review, *BMC Pulmonary Medicine*, 2009;9(1):24.

17. Thorpe KE and Howard DH, The rise in spending among Medicare beneficiaries: the role of chronic disease prevalence and changes in treatment intensity, *Health Affairs*, 2006;25(5):w378.

18. Ibid..

19. Hornbrook M, Hurtado A, et al., Health care episodes: definition, measurement and use, *Medical Care Research and Review*, 1985;42(2):163.

20. Hornbrook, et al.

21. Ibid.

22. Ibid.

23. Pham HH, Schrag D, et al., Care patterns in Medicare and their implications for pay for performance, *New Eng J Med*, 2007;356(11): 1130-1139.

24. Hogan C, Lunney J, et al., Medicare beneficiaries' costs of care in the last year of life, *Health Affairs*, 2001;20(4):188.

25. Tompkins CP, Higgins AR, et al., Measuring outcomes and efficiency in Medicare value-based purchasing, *Health Affairs*, 2009;28(2):w251.

26. Chung J, Kaleba E, Wozniak G, A Framework for Measuring Healthcare Efficiency and Value. PCPI Work Group on Efficiency and Cost of Care. Aug 2008. White Paper.

27. Hussey PS, De Vries H, et al., A systematic review of health care efficiency measures, *Health Services Research*, 2009;44(3):784-805.

## APPENDIX A—SPECIFICATIONS FOR COST AND RESOURCE USE MEASURES 2011 (Cycle 2)

The following tables present the detailed measure specifications for the recommended consensus standards. All information summarized 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.

**Pulmonary**



**Bone/Joint**



| | 1560: Relative Resource Use for People with Asthma |
|---|---|
| Steward | NCQA |
| Description | The risk-adjusted relative resource use by health plan members with asthma during the measurement year. |
| Resource Use Measure Type | Per capita (population- or patient-based) |
| Data Source | Administrative claims<br>Electronic Clinical Data : Electronic Health Record, Imaging/Diagnostic Study, Laboratory, Pharmacy<br>Paper Records |
| Level of Analysis | Clinician : Group/Practice<br>Health Plan<br>Integrated Delivery System<br>Population : National, Regional |
| Clinical Framework Description | 2 Eligibility Criteria: An encounter with a diagnosis; or by multiple asthma medication events. An organization must use both methods to identify the eligible population, but a member only needs to be identified by one to be included in the measure. To identify the eligible population for measurement:<br>Step 1: Health Plan members are identified as having persistent asthma by meeting at least one of the following criteria during both the measurement year and the year prior to the measurement year. Criteria need not be the same across years.<br>• At least one ED visit (Tables ASM-A and ASM-B) with asthma as the principal diagnosis, or<br>• At least one acute inpatient claim/encounter (Tables ASM-B) with asthma as the principal diagnosis (Table ASM-A), or<br>• At least four outpatient asthma visits (Table ASM-B) with asthma as one of the listed diagnoses (Table |

| | 1560: Relative Resource Use for People with Asthma |
|---|---|
| | ASM-A) and at least two asthma medication dispensing events (Table ASM-C), or<br>• At least four asthma medication dispensing events (Table ASM-C)<br>Step 2: Since a member can be identified as having persistent asthma using only leukotriene modifiers as the sole asthma medication dispensed in that year, these members must also have at least one diagnosis of asthma (Table ASM-A), in any setting, in the same year as the leukotriene modifier prescription (e.g. measurement year or year prior to the measurement year).<br><br>Exclusions:<br>1) Active cancer. Exclude members who had at least one face-to-face encounter, in any setting, with any diagnosis of cancer in conjunction with any treatment code (Table RRU-A), during the measurement year.<br>2) ESRD. Exclude members who had at least one face-to-face encounter with any code to identify ESRD (Table RRU-B), during the measurement year.<br>3) Organ transplant. Exclude members who had at least one face-to-face encounter, in any setting, with any code to identify organ transplant (Table RRU-C), during the measurement year.<br>4) HIV/AIDS. Exclude members who had at least two face-to-face encounters in an outpatient or nonacute inpatient setting, or at least one face-to-face encounter in an acute inpatient or ED setting, with any diagnosis of HIV (Table RRU-D), with different dates of service during the measurement year. Refer to Table RRU-E for codes to identify visit type.<br>5) Members diagnosed with emphysema, COPD, cystic fibrosis or acute respiratory failure (Table ASM-E) on or prior to December 31 of the measurement year. |
| 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. The standard pricing approach is based on the following sources of data:<br>• Relative values from the Medicare Fee Schedule (Resource-Based Relative Value Scale, or RBRVS)<br>• Pharmacy prices published by First Bank Data<br>• Inpatient prices based on a model that uses a broad set of averages, representing different local, regional and national health plans across the country.<br>A plan maps a standard price to each service, multiplies it by the number of units of service and reports the resulting standard cost. It then calculates total standard costs for eligible members across different areas of clinical care and aggregates standard costs across services and members to compute the overall relative resource use.<br>All RRU measures report the standard cost for the following categories.<br>o Inpatient Facility<br>o Surgery and Procedure<br>o Inpatient Services<br>o Outpatient Services<br>o Evaluation and Management (E&M)<br>o Inpatient Services<br>o Outpatient Services<br>o Diagnostic Laboratory Services<br>o Diagnostic Imaging Services<br>o Pharmacy, Ambulatory |
| Tested Population | Commercial; Medicaid |
| Resource Use Service Categories | Inpatient services: Inpatient facility services; Evaluation and management; Procedures and surgeries; Imaging and diagnostic; Lab services; Admissions/discharges<br><br>Ambulatory services: Outpatient facility services; Emergency Department; Pharmacy; Evaluation and management; Procedures and surgeries; Imaging and diagnostic; Lab services |

| | 1560: Relative Resource Use for People with Asthma |
|---|---|
| Attribution Approach | Specifications: 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. |
| Risk Adjustment | The current risk model utilized by NCQA is based on components of the CMS-HCC risk adjustment methodology and accounts for age, gender, and HHC-RRU risk classifications that predict cost variability. For each condition, members are assigned to a clinical cohort category that provides a more specific classification of the condition. A members age, gender, and HCC category determines their risk score (cohort). NCQA then calculates the average per-member per-month (PMPM) cost for each cohort then weights that cost by the total member months within each cohort. Each plan will have its own weight for each cohort since case-mix varies across plans. These weighted cohort PMPMs are then summed across all cohorts to estimate total resource use that would be expected if the "average" plan had the same case-mix as the plan in question. The ratio of the observed- to-expected PMPM utilization indicates the degree to which a plan deviates from expected performance. This is known as indirect standardization. |
| Stratification | NCQA collects resource measures at the plan level and summarizes across reporting cohorts along the following dimensions:<br>Product line (3 levels): Commercial, Medicaid, and Medicare;<br>Reporting type (2 levels): HMO and PPO;<br>Area level (2 levels): national and region;<br>Resource use or utilization (11 levels): inpatient facility, procedure and surgery (inpatient and outpatient), evaluation and management (inpatient and outpatient), laboratory services, imaging services, ambulatory pharmacy, inpatient discharges, emergency department discharges.<br>Stratification of RRU results to control for individual confounding variables is not performed since age, gender and risk variables (comorbidity and disease interactions) that affect healthcare costs are accounted for in the RRU-HCC risk adjustment process. These include age and gender along with one of the 13 assigned HCC-RRU risk categories (e.g. male 18-44 HCC-RRU 1; male 18-44 HCC-RRU 2; male 18-44 HCC-RRU 3; etc…). |

1412

| | 1561: Relative Resource Use for People with COPD |
|---|---|
| Steward | NCQA |
| Description | The risk-adjusted relative resource use by health plan members with COPD during the measurement year. |
| Resource Use Measure Type | Per capita (population- or patient-based) |
| Data Source | Administrative claims<br>Electronic Clinical Data: Electronic Health Record, Imaging/Diagnostic Study, Laboratory, Pharmacy<br>Paper Records |
| Level of Analysis | Clinician : Group/Practice<br>Health Plan, Integrated Delivery System,<br>Population : Community, National, Regional |
| Clinical Framework Description | Members are identified for the eligible population of the measure with a diagnosis of COPD (Table SPR-A) present anytime during the measurement year and who were continuously enrolled for a two year period (the measurement year and the year prior).<br>Codes to Identify COPD:<br>Chronic bronchitis-ICD-9 Diagnosis: 491<br>Emphysema -ICD-9 Diagnosis: 492<br>COPD -ICD-9 Diagnosis: 496<br><br>Exclusions:<br>1) Active cancer. Exclude members who had at least one face-to-face encounter, in any setting, with any diagnosis of cancer in conjunction with any treatment code (Table RRU-A), during the measurement year. |

| | 1561: Relative Resource Use for People with COPD |
|---|---|
| | 2) ESRD. Exclude members who had at least one face-to-face encounter with any code to identify ESRD (Table RRU-B), during the measurement year.<br>3) Organ transplant. Exclude members who had at least one face-to-face encounter, in any setting, with any code to identify organ transplant (Table RRU-C), during the measurement year.<br>4) HIV/AIDS. Exclude members who had at least two face-to-face encounters in an outpatient or nonacute inpatient setting, or at least one face-to-face encounter in an acute inpatient or ED setting, with any diagnosis of HIV (Table RRU-D), with different dates of service during the measurement year. Refer to Table RRU-E for codes to identify visit type.<br>5) Members diagnosed with emphysema, COPD, cystic fibrosis or acute respiratory failure (Table ASM-E) on or prior to December 31 of the measurement year. |
| 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. The standard pricing approach is based on the following sources of data:<br>• Relative values from the Medicare Fee Schedule (Resource-Based Relative Value Scale, or RBRVS)<br>• Pharmacy prices published by First Bank Data<br>• Inpatient prices based on a model that uses a broad set of averages, representing different local, regional and national health plans across the country.<br>A plan maps a standard price to each service, multiplies it by the number of units of service and reports the resulting standard cost. It then calculates total standard costs for eligible members across different areas of clinical care and aggregates standard costs across services and members to compute the overall relative resource use.<br>All RRU measures report the standard cost for the following categories.<br>o Inpatient Facility<br>o Surgery and Procedure<br>o Inpatient Services<br>o Outpatient Services<br>o Evaluation and Management (E&M)<br>o Inpatient Services<br>o Outpatient Services<br>o Diagnostic Laboratory Services<br>o Diagnostic Imaging Services<br>o Pharmacy, Ambulatory |
| Tested Population | Commercial; Medicaid; Medicare |
| Resource Use Service Categories | Inpatient services: Inpatient facility services, Evaluation and management, Procedures and surgeries, Imaging and diagnostic, Lab services<br>Admissions/discharges<br>Ambulatory services: Outpatient facility services, Emergency Department; Pharmacy, Evaluation and management, Procedures and surgeries, Imaging and diagnostic, Lab services |
| Attribution Approach | Specifications: 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. |
| Risk Adjustment | The current risk model utilized by NCQA is based on components of the CMS-HCC risk adjustment methodology and accounts for age, gender, and HHC-RRU risk classifications that predict cost variability. For each condition, members are assigned to a clinical cohort category that provides a more specific classification of the condition. A members age, gender, and HCC category determines their risk score (cohort). NCQA then calculates the average per-member per-month (PMPM) cost for each cohort then weights that cost by the total member months within each cohort. Each plan will have its own weight for each |

| | 1561: Relative Resource Use for People with COPD |
|---|---|
| | cohort since case-mix varies across plans. These weighted cohort PMPMs are then summed across all cohorts to estimate total resource use that would be expected if the "average" plan had the same case-mix as the plan in question. The ratio of the observed- to-expected PMPM utilization indicates the degree to which a plan deviates from expected performance. This is known as indirect standardization. |
| Stratification | NCQA collects resource measures at the plan level and summarizes across reporting cohorts along the following dimensions:<br>Product line (3 levels): Commercial, Medicaid, and Medicare;<br>Reporting type (2 levels): HMO and PPO;<br>Area level (2 levels): national and region;<br>Resource use or utilization (11 levels): inpatient facility, procedure and surgery (inpatient and outpatient), evaluation and management (inpatient and outpatient), laboratory services, imaging services, ambulatory pharmacy, inpatient discharges, emergency department discharges.<br>Stratification of RRU results to control for individual confounding variables is not performed since age, gender and risk variables (comorbidity and disease interactions) that affect healthcare costs are accounted for in the RRU-HCC risk adjustment process. These include age and gender along with one of the 13 assigned HCC-RRU risk categories (e.g. male 18-44 HCC-RRU 1; male 18-44 HCC-RRU 2; male 18-44 HCC-RRU 3; etc…). |

1413
1414

| | 1609: ETG Based hip/knee replacement cost of care measure |
|---|---|
| Steward | Ingenix |
| Description | This submission is for Hip/Knee Replacement procedure episodes and will cover both measures at the Hip Replacement and Knee Replacement PEGs. The measure focuses on resources used to deliver episodes of care for patients who have undergone a hip or knee replacement and assigns a level of severity (e.g., resources per episode for Knee Replacement, severity level 1 episodes). Hip Replacement and Knee Replacement episodes are initially defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating the condition. The Procedure Episode Group (PEG) methodology uses the ETG results and further logic to creating a procedure episode that focuses on the Hip Replacement and Knee Replacement component of the care. Procedure episodes identify a unique procedure event as well as the related services performed before and after the procedure including workup and therapy prior to the procedure as well as post-op activities such as repeated surgery and patient follow-up. Together, the ETG and PEG methodologies identify the services involved in diagnosing, managing and treating patients with Hip/Knee Replacements. A methodology to assign a severity level to each episode is employed to group Hip and Knee Replacement episodes by level of risk.<br><br>Multiple types of resources can be measured for Hip/Knee Replacement episodes, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per episode and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons. |
| Resource Use Measure Type | Per episode |
| Data Source | Administrative claims |
| Level of Analysis | Clinician : Group/Practice, Individual, Team, Facility, Health Plan, Integrated Delivery System<br>Population : Community, County or City, National, Regional, State |
| Clinical Framework Description | This measure identifies patients with Hip/Knee Replacement and creates Hip/Knee Replacement episodes of care using the ETG and PEG methodologies described in the ETG_PEG Construction Logic attached in our response to S.2. Each procedure episode of Hip/Knee Replacement is characterized by a PEG Anchor Category ID that specifies the type of procedure; the PEG Anchor Category ID representing Hip Replacement |

| | 1609: ETG Based hip/knee replacement cost of care measure |
|---|---|
| | is 71518 and the PEG Anchor Category ID representing Knee Replacement is 71918.<br><br>An ETG/PEG episode of Hip/Knee Replacement will contain all clinically relevant information related to the procedure. The Hip/Knee Replacement episode clinical framework is defined by the services, or claim lines, that can begin an episode, the primary and incidental diagnosis relationships involved and how records group to an episode, including relative strength of relationship. |
| Costing Method | The financial amounts used should be complete and valid, reflecting the total payments related to the service. The financial amount used in resource measurement should reflect all payments for a service, including those made to the provider by payer, patient and other entities. Allowed payments will reflect both the quantity of different services provided as well as the actual unit price of those same services. |
| Tested Population | Commercial |
| Resource Use Service Categories | Inpatient services: Inpatient facility services, Admissions/discharges<br>Ambulatory services: Outpatient facility services, Emergency Department, Pharmacy, Evaluation and management, Procedures and surgeries, Imaging and diagnostic, Lab services |
| Attribution Approach | Guidelines: For physician measurement, the primary surgeon is typically attributed the episode, although applications of attribution could be developed to support an alternate approach. Both activity-based and population-based approaches should be supported. As a guideline, four different general options for physician episode attribution can be considered to attribute episodes to individual providers – three activity-based and one population-based approach.<br>Approach 1 - Physician Episode Attribution using Professional Service Costs. This attribution approach identifies the responsible physician for an episode as that provider rendering the greatest amount of professional service costs during the episode.<br>Approach 2 - Physician Episode Attribution using Episode Clusters. This attribution approach identifies the responsible physician for an episode as that provider in the peer group owning the greatest number of "clusters" within the episode.<br>Approach 3 - Physician Episode Attribution using Non-Acute Evaluation and Management (E/M) Visits. This attribution approach identifies the responsible physician for an episode as that physician providing the greatest number of non-acute E/M visits within the episode.<br>Approach 4 - Physician Episode Attribution using a Primary Care, Population-based Approach. This approach requires two important steps: 1) Identification of a PCP for each member. 2) Identify the patient's assigned PCP during the episode period. |
| Risk Adjustment | The level of severity assigned to an episode is used to support risk adjustment. The risk adjustment approach includes three important steps:<br>1. Compute the observed experience for the provider being measured, across all episodes to be included in the comparison;<br>2. Compute the experience for peers or a best practice benchmark. Compute this experience at the level of the risk adjustment, in this case base procedure (hip or knee replacement) and severity level. For a peers benchmark, average cost per episode across all peers for the base procedure and severity level can be computed;<br>3. Compare the observed experience with the risk adjusted peers or benchmark experience – often called the "expected" result. This expected result is adjusted to reflect both the peers/benchmark levels of performance and also the provider's own case mix of episodes by condition and level of severity. The ratio of observed to expected results can be termed the relative cost ratio and is a risk adjusted measure. |
| Stratification | The severity level can then be used to stratify episodes by severity, measured as resource consumption. |

1415
1416

| | 1611: ETG Based Pneumonia cost of care measure |
|---|---|
| Steward | Ingenix |
| Description | The measure focuses on resources used to deliver episodes of care for patients with pneumonia. Pneumonia episodes are defined using the Episode Treatment Groups (ETG) methodology and describe the unique presence of the condition for a patient and the services involved in diagnosing, managing and treating pneumonia. A number of resource use measures are defined for pneumonia episodes, including overall cost of care, cost of care by type of service, and the utilization of specific types of services. Each resource use measure is expressed as a cost or a utilization count per episode and comparisons with internal and external benchmarks are made using risk adjustment to support valid comparisons. |
| Resource Use Measure Type | Per episode |
| Data Source | Administrative claims, Other: Both medical and pharmacy administrative service records (claims or encounters) are used to support the measures. Member enrollment span, pharmacy benefit status and age and gender are also required. Provider characteristics, including specialty and unique provider identifier also have importance to support episode grouping, attribution and definition of peers. |
| Level of Analysis | Clinician : Group/Practice, Individual, Team<br>Facility<br>Health Plan<br>Integrated Delivery System<br>Population: Community, County or City, National, Regional, State |
| Clinical Framework Description | The pneumonia measure's episodes are defined using the Episode Treatment Group (ETG) methodology. The pneumonia ETG episode building process that supports pneumonia resource use measures has four important steps:<br>Step 1: Identify Records; Assign Record Type and Anchor Records, Classify Diagnoses and Procedures<br>Step 2: Build Episodes from Anchor Records<br>Step 3: Group Non-Anchor Records to Episodes<br>Step 4: Finalize the Episodes (identify co-morbidities and complicating factors, and assign episode severity) |
| Costing Method | The financial amounts used should be complete and valid, reflecting the total payments related to the service. The financial amount used in resource measurement should reflect all payments for a service, including those made to the provider by payer, patient and other entities. Allowed payments will reflect both the quantity of different services provided as well as the actual unit price of those same services. |
| Tested Population | Commercial |
| Resource Use Service Categories | Inpatient services: Inpatient facility 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 | Guidelines: Both activity-based and population-based approaches should be supported. As a guideline, four different general options for physician episode attribution can be considered to attribute episodes to individual providers – three activity-based and one population-based approach.<br>Approach 1 - Physician Episode Attribution using Professional Service Costs. This attribution approach identifies the responsible physician for an episode as that provider rendering the greatest amount of professional service costs during the episode.<br>Approach 2 - Physician Episode Attribution using Episode Clusters. This attribution approach identifies the responsible physician for an episode as that provider in the peer group owning the greatest number of "clusters" within the episode.<br>Approach 3 - Physician Episode Attribution using Non-Acute Evaluation and Management (E/M) Visits. This attribution approach identifies the responsible physician for an episode as that physician providing the greatest number of non-acute E/M visits within the episode. |

| | 1611: ETG Based Pneumonia cost of care measure |
|---|---|
| | Approach 4 - Physician Episode Attribution using a Primary Care, Population-based Approach. This approach requires two important steps: 1) Identification of a PCP for each member. 2) Identify the patient's assigned PCP during the episode period. |
| Risk Adjustment | ETG first assesses the observed co-morbidities and condition status factors for an episode and the patient's age and gender. ETG then assigns a weight to each factor found to influence the relative risk of an episode of pneumonia. These weights and factors are condition-specific and were estimated using pneumonia episode results for a large population. The overall severity score for an episode is the sum of these weights for all factors observed. Using the severity score, a severity level is created, with each pneumonia episode assigned to one of four severity levels. The level of severity assigned to an episode is used to support risk adjustment. The risk adjustment approach includes three important steps:<br>1. Compute the observed experience for the provider being measured, across all episodes to be included in the comparison;<br>2. Compute the experience for peers or a best practice benchmark. Compute this experience at the level of the risk adjustment, in this case base procedure (hip or knee replacement) and severity level. For a peers benchmark, average cost per episode across all peers for the base procedure and severity level can be computed;<br>3. Compare the observed experience with the risk adjusted peers or benchmark experience – often called the "expected" result. This expected result is adjusted to reflect both the peers/benchmark levels of performance and also the provider's own case mix of episodes by condition and level of severity. The ratio of observed to expected results can be termed the relative cost ratio and is a risk adjusted measure. |
| Stratification | ETG stratifies episodes by the intensity of service, or total cost. For a given episode, a severity score is assigned based on demographic factors (gender and age) and the presence of comorbidities and complications. Once a severity score is determined, a severity level, a number between 1 and 4 is assigned based on a table that relates severity levels to severity scores for each ETG. The severity level can then be used to stratify episodes by severity, measured as resource consumption. |

## APPENDIX B—STEERING 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—TECHNICAL ADVISORY PANELS

### 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

**Pulmonary Technical Advisory Panel**

**Kurtis Elward, MD, MPH (Co-Chair)**
Family Medicine of Albermarle, Charlottesville, VA

**Janet Maurer, MD, MBA (Co-Chair)**
American College of Chest Physicians, Northbrook, IL

**Gerene Bauldoff, PhD, RN**
The Ohio State University, School of Nursing, Columbus, OH

**Kathryn Blake, PharmD**
Nemours Children's Clinic, Jacksonville, FL

**Dale Bratzler, DO, MPH**
University of Oklahoma, Health Sciences Center, Oklahoma City, OK

**Zab Mosenifar, MD**
Cedars Sinai Medical Center, Los Angeles, CA

**Linus Santo Tomas, MD, MS**
Pulmonary & Critical Care, Medical College of Wisconsin, Milwaukee, WI

**Michael Schatz, MD, MS**
Kaiser Permanente, Oakland, CA

**Richard Stanford, PharmD, MS**
GlaxoSmithKline, Research Triangle Park, NC

**Bone/Joint Technical Advisory Panel**

**James Weinstein, DO, MS(Chair)**
The Dartmouth Institute for Health Policy; Dartmouth-Hitch Clinic, Lebanon, NH

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

**Elizabeth Paxton, MA**
Kaiser Permanente, Oakland, CA

**John Ratliff, MD, FACS**
Thomas Jefferson University, Philadelphia, PA

**Catherine Roberts, MD**
Mayo Clinic, Phoenix, AZ

**Craig Rubin, MD**
University of Texas Southwestern Medical School, Dallas, TX

**Patricia Sinnott, PT, PhD, MPH**
VA health Economics Resource Center, Menlo Park, CA

**Cancer Technical Advisory Panel**

**David Penson, MD, MPH (Chair)**
Vanderbilt University Medical Center, Nashville, TN

**Rohit Borker, PhD**
GlaxoSmithKline, Philadelphia, PA

**Steven Chen, MD, MBA**
California Medical Association, Camerillo, CA

**Timothy Gilligan, MD**
Cleveland Clinic Taussig Cancer Institute, Cleveland, OH

**Stephen Grossbart, PHD**
Catholic Healthcare Partners, Cincinnati, OH

**Dwight Kloth, PharmD**
Fox Chase Cancer Center, Philadelphia, PA

**Louis Potters, MD, FACR**
North Shore-Long Island Jewish Health System, New Hyde Park, NY

**Jay Schukman, MD**
Anthem Blue Cross and Blue Shield, Richmond, VA

**John Skibber, MD**
University of Texas-MD Anderson Cancer Center, Houston, TX

**Louise Walter, MD**
University of California - San Francisco, San Francisco, CA

## 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.