NATIONAL QUALITY FORUM

TO: Consensus Standards Approval Committee (CSAC)

FR: Taroon Amin, Senior Director
Ashlie Wilbon, Senior Project Manager
Evan M. Williamson, Project Analyst

RE: Result of Voting for *National Voluntary Consensus Standards for Cost and Resource Use Measures (Cycle 2)*

DA: February 24, 2012

CSAC ACTION REQUIRED

Pursuant to the National Quality Forum's (NQF's) version 1.9 of the Consensus Development Process (CDP), the CSAC may consider approval of five candidate consensus standards as specified in the "voting draft" of *National Voluntary Consensus Standards for Cost and Resource Use (Cycle 2).*

Cycle 2 Measures Recommended for Endorsement:

- (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)

Additionally, for one measure the Committee was unable to reach consensus (split vote):

- (1595) ETG based diabetes cost of care measure (Ingenix)

Accompanying this memo are the following documents:

1. **Cycle 2 Resource Use Draft Report.** The voting draft report has been updated to reflect the changes made following Steering Committee discussion of public and member comments. The complete draft report and supplemental materials are available on the project page.
2. **Comment table for Cycle 2 Resource Use Draft Report.** While staff has identified themes within the comments received, all comments did not fit within the themes. This table lists all 87 comments received and the responses.

BACKGROUND

In the context of this project, resource use measures are defined as broadly applicable and comparable measures of health services counts that are applied to a population or event, counting the frequency of defined health system resources. 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.

This CDP project is the second phase of a two-phase effort. Phase one, which began in 2009, was aimed at understanding resource use measures and identifying the important attributes to consider in their evaluation. During this phase, the current NQF Measure Evaluation Criteria used for the evaluation of quality measures was reviewed and refined by the Resource Use Steering Committee to address the unique aspects of resource use measures, resulting in the NQF Resource Use Measure Evaluation Criteria. These criteria were reviewed and approved by CSAC in November 2010.

Phase two was divided into two measure review cycles between which fourteen condition areas were identified. A single Steering Committee was used across both phases of work, with an additional four Technical Advisory Panels (TAPs) in phase two to assist the Committee in evaluating the measures' clinical and methodological aspects. Information in this memo and the Cycle 2 Resource Use Draft Report reflect the discussion and overarching issues the Committee identified while evaluating cost and resource use measures submitted to the project.

DRAFT REPORT

Building upon the Cycle 1 report, the Cycle 2 Resource Use Draft Report presents the results of the evaluation of eleven measures considered under the National Quality Forum's CDP. Four are 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)

One measure had a split vote by the Steering Committee:

- (1595) ETG based diabetes cost of care measure (Ingenix)

Six measures were reviewed and not recommended for endorsement:

- (1591) ETG-based congestive heart failure (CHF) cost of care measure (Ingenix)
- (1594) ETG-based coronary artery disease (CAD) cost of care measure (Ingenix)
- (1599) ETG-based non-condition specific cost of care measure (Ingenix)
- (1603) ETG-based hip fracture cost of care measure (Ingenix)
- (1605) ETG-based asthma cost of care measure (Ingenix)
- (1608) ETG-based chronic obstructive pulmonary disease cost of care measure (Ingenix)

The TAPs and Committee encountered several overarching issues during their discussions and evaluations of the measures. Some issues varied by developer as each developer submitted

measures with very distinct approaches. The Committee factored these issues into their ratings and recommendations for multiple measures, recognizing the need to balance the quantity and specificity of information required to adequately evaluate the measure and the burden on the developer to provide this information.

COMMENTS AND THEIR DISPOSITION

NQF received 87 comments from 11 organizations and individuals on measures both recommended and not recommended for endorsement as well as general comments. The distribution of individual comments by Member Council follows:

- Consumers: 23 comments
- Health Professionals: 8 comments
- Purchasers: 49 comments
- Public Health/Community: 0 comments
- Health Plans: 6 comment
- Quality Measurement, Research, and Improvement: 0 comments
- Providers: 1 comments
- Supplier and Industry: 0 comment
- Non-members: 0 comments

To view the individual comments, refer to the comment table with the responses to each comment and any actions taken by the Steering Committee and/or measure developers.

Due to the volume and repetition of comments, the major themes of the comments and issues were identified for Committee discussion. There were nine major themes identified, many of which overlapped with the themes identified during the cycle 1 comment period. The Committee discussed and provided responses to each theme.

General Comments: Major Themes/Issues

1. Application of costing approaches
2. Splitting costing approaches into separate measures
3. Higher bar for resource use measure evaluation (than for quality measures)
4. Measures in use should be endorsed
5. Complexity of the resource use measures from an episode grouper
6. Implementation costs associated with Ingenix measures
7. Risk adjustment model
8. Preference for specifications compared to guidelines
9. Burden of validity testing

Comment Themes and Committee Responses

Theme 1- Application of Costing Approaches

Description. Comments submitted expressed strong views on the usefulness of cost measures based on actual prices paid for comparison of prices in markets nationally. Commenters argued that measures using actual prices should be paired with measures using standardized prices to better understand market influence and the margin between prices paid and resource use.

> *Committee Response*: 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 within regions and markets. The Committee agreed that both approaches could be appropriate for different applications. However, the Committee's decision to recommend (or not recommend) individual measures should not be interpreted as driven by simply the measure's costing approach. A measure-by-measure decision was made on the appropriateness of the costing approach given other measure characteristics, resulting in the endorsement of both types of measures. Reliability and validity was examined through the interaction of the measure's specified level of measurement, risk adjustment model, and other measure characteristics. There was agreement that actual prices paid by health plans to individual clinicians is important to measure and report; for example, regional comparisons at the individual clinician level where environmental factors may not be as prominent, or nationally at higher levels of measurement (i.e., health plan level). The Committee did, however, express concern over applying an actual price approach for national comparisons at an individual clinician level. Specifically, the Committee noted the potential for misinterpreting clinician resource use in national reporting. This pricing approach includes environmental factors (i.e., local facility and wage index) that may be outside of an individual clinician's control. The Committee agreed that when actual prices paid are reported, utilization counts should be reported as well.

Theme 2- Splitting costing approaches into separate measures

Description. Comments submitted questioned the need to separate costing approaches into separate measures, arguing the need for both approaches to be included in one measure.

> *Committee Response*: The Committee agreed early in the evaluation process that a single measure should allow for only one costing approach (actual prices paid or standardized pricing) to ensure consistent and accurate comparisons of measure results. For use as a national consensus standard, measure results should unambiguously reflect differences in performance for an accountable entity, not differences in the type of data an entity choses to submit (actual prices or standardized prices). As such, developers that allowed for user flexibility in the costing approach were asked to split their measures into two separate measures where only one approach is specified in a single measure. Endorsing measures with a single costing approach does not preclude the use of both measures as a pair. Developers also had the option to select a single costing approach to be applied to the measure. Health Partners elected to split their measure into two, while Ingenix selected to have one measure based on actual price paid reviewed.

Theme 3- Higher bar for resource use measure evaluation

Description: Commenters expressed concern that the report appeared to describe an evaluation standard that was higher than that used for quality measures, arguing that the evaluation of resource use measures should be held to the same standard as quality measures.

Committee Response: The resource use measure evaluation criteria are the same criteria used for quality measures; specifically, importance to measure and report, scientific acceptability of the measure properties, usability and feasibility. In order to customize the evaluation to specific components in resource use measures, the Steering Committee, in its first phase of work, sought to identify how resource use measures should be specified, and how to evaluate reliability and validity in these types of measures. The result of this effort is the NQF resource use measure evaluation criteria and the resource use specification modules.

The Committee identified five "modules" to describe the way resource use measures should be specified including data protocol, clinical logic, construction logic, adjustments for comparability, and reporting. The modules sought to provide developers with a familiar framework in which resource use measures are often constructed. The submission process was mirrored after the modules and vetted by most developers who submitted measures to the project (including Ingenix and NCQA).

While some of the measure evaluation sub-criteria needed to be adapted for resource use, including the importance and usability subcriteria; the remaining criteria remained unchanged from the criteria that are applied to quality measures. When evaluating the measures, the Committee applied the same criteria to all submitted measures in the same manner while taking into consideration some of the unique constructs of resource use measures and the nature of the interactive components of the specifications.

Both quality and resource use measures must demonstrate adequate reliability and validity testing at the lowest specified level of analysis. The Committee's determination of adequate testing and results relied on expert judgment of an independent statistic consultant, the Technical Advisory Panels, and members of the Steering Committee to consider: (1) if the developers testing was appropriate for the specified measure; (2) if the scope of testing including the representiveness and sample size was adequate for the specified level of analysis; and (3) if the results indicated an acceptable level of reliability and validity. This standard is consistent across both types of measures.

Theme 4- Measures in use should be endorsed

Description: Commenters argue that measures that are already widely in use should meet the field testing requirements and this should be taken into consideration when making recommendations for endorsement. Because a measure is in use, it is inherently usable and feasible.

Committee Response: The Committee acknowledged that resource use measures have been in use in the commercial/private sector for many years, but have not been subject to the review and scrutiny that most quality measures have. In addition to the various complex methods and approaches for measuring the same types of costs/resources, there

is limited published peer reviewed literature about the reliability and validity of these measures. This effort marks the first time that many of these measures have been subject to a systematic review of the methodology and scientific acceptability. As such, the wide use of these measurement approaches does not inherently imply that quality or resource use measures are acceptable for endorsement. The Committee also acknowledges the sensitive nature of some measures where financial investments have been made on behalf of purchasers and other users for reporting and understanding costs/resource use. The context and process by which measures become endorsed as NQF standards requires that the measures meet each of the four criteria and qualify for use for accountability and quality improvement purposes. While the current use of the measures is taken into consideration (within the usability criteria) by the Committee during evaluation, it does not imply the measure meets the criteria for endorsement nor does it satisfy the scientific acceptability criteria.

Theme 5- Complexity of the Resource Use measures from an episode grouper

Description: Commenters expressed concern that measures submitted by Ingenix were not endorsed due to their complexity. They argue that resource use and cost measures that use an episode grouper are inherently complex. Alternatively, Commenters also believe that due to the complexity of these measures they should be examined before the typical three year review cycle. This shorter cycle for updating these measures will help to solicit feedback from the field on the implementation process of these measures.

Committee Response: The Committee recognizes that resource use measures, including those derived from episode groupers are inherently complex. This complexity should not, however, hinder the transparency, clarity, and ability to deconstruct the measure for understanding. Further, the Committee chose to recommend measures based on individual measure characteristics, rather than disregarding any measure due to its inherent complexity. The Committee noted that the ERG risk adjuster is very complex and still endorsed several measures. The Committee agreed that resource use measures should be held to the same standard as quality measures, and evaluated against the same criteria; specifically, importance to measure and report, scientific acceptability of the measure properties, usability and feasibility. NQF will strongly consider a shorter cycle for updating these measures considering the concerns raised.

Theme 6- Cost of the measures submitted by Ingenix

Description: One commenter believed very strongly that the Committee should acknowledge the widespread use of Ingenix measures even in light of their costs. While another commenter expressed concern over the cost of the Ingenix measures, include cost of ETGs, ERGs, PEGs and the cost of implementation.

Committee Response: The Committee considered the cost of the Ingenix product (ETGs, ERGs, PEGs) in the feasibility criterion of the measure evaluation as indicated by the policy on endorsement of proprietary performance measures. This policy is not unique to resource use measures and is applied in the evaluation of proprietary quality measures with fees as well. While some users may find the cost of the episode grouper reasonable, the use of these measurements does not inherently imply the measures are acceptable for

endorsement. The issue of the cost of the measures submitted by Ingenix was weighted differently by the various stakeholders represented in the Steering Committee. The Committee also weighed the potential burden these costs may carry if these measures were adopted for regional or national reporting programs requiring that organizations take on these costs to participate. The Committee agreed that while the issue of cost was taken into consideration, it was not a deciding factor in the recommendations for any of the measures.

Theme 7-Risk adjustment model

Description: Commenters disagreed that factors in the risk adjustment model and severity model should be confirmed to be a contributor to the outcome of the measure. One commenter was very concerned that the Committee was too focused on the scientific validity and that the variables used in the risk adjustment methods were actually correlated with outcomes (as well as clinically significant).

Committee Response: The Committee looked to guidance provided by the measure evaluation criteria and the NQF Measure Testing and Evaluation Scientific Acceptability of Measure Properties Task Force report. For resource use measures and quality measures, an evidence-based risk adjustment strategy (e.g., risk models, risk stratifications) should be based on patient clinical factors that influence the measured outcome (page 24). When evaluating the validity testing of the measure, the Committee sought to ensure that the data and sample used for development and validation are reflective of its intended population. The Committee agreed that measure developers have a responsibility to demonstrate quantitatively, the relative contribution of risk factors, risk model performance metrics and the assessment of adequacy in the context of norms for risk models. The Committee argued that these testing requests are similar and aligned with quality measures.

Theme 8-Preference for specification compared to guidelines

Description: Commenters believed that the Steering Committee favored specifications over guidelines. The concerns specifically referenced Emerging Principle 1 favoring specifications for the resource use measure construct.

Committee Response: The Committee did not express preference for specifications or guidelines. The submission process required that the measure clinical logic, construction logic, and adjustments for comparability details be submitted as specifications; however, all submission items within the data protocol and reporting modules allowed for flexibility. The measure submission was intentionally designed with this flexibility in these modules of the measure.

Theme 9- Burden of validity testing

Description: Commenters expressed concern that the validity testing requirements are overly prescriptive and should not require a chart review as a necessary validity check. Chart reviews are expensive and are also susceptible to deficiencies that limit the accuracy of data extraction.

Committee Response: The Committee agreed that adequate validity testing is required for resource use measures in addition to quality measures, relying on guidance from the NQF Measure Testing and Evaluation Scientific Acceptability of Measure Properties Task Force report. Validity testing can be done at the data element or the measure score level. If the developers choose to demonstrate data element validity, patient-level information on individual patients (e.g., count of medication provided) should demonstrate that the data elements are correct and the correctly identify differences in resource use (page 14; page 31). However, data element validity does not need to be conducted for every single data element. Testing can include only those critical data elements. Developers also have the option of measure score validity testing where developers can demonstrate correlation of measure score results with another valid indicator of resource use. Developers have the responsibility to demonstrate the data elements and/or measure score are reliable and valid in their testing. Emerging principle 7 should not be interpreted as chart reviews are a necessary validity check, but rather, when demonstrating validity of data elements they should be evaluated against an authoritative source (e.g., a similar measure that has been validated, a validated tool). The Committee further stated that during the measure evaluation, distinguishing between the two testing approaches (score or data element level) was not a major discussion or concern for any of the measures.

Measure Specific Comments on Recommended Measures

(1560) Relative Resource Use for People with Asthma (NCQA)
(1561) Relative Resource Use for People with COPD (NCQA)
Description: Comments received for the two NCQA measures were similar. Commenters disagreed with the Committee's request for sample size requirements of 400 for NCQA measures. They argue that sample size requirements are overly restrictive and measure developers should have enough sample size to demonstrate reliability of 0.7. Moreover, commenters were concerned about this measure's use of administrative data as they are notoriously inaccurate, implementation of the measure may be overly burdensome, and problems with the use of diagnostic codes to distinguish between asthma and COPD in older persons. Commenters encourage the developers to use historical data to confirm and distinguish between COPD and asthma.

Committee Response: The Committee evaluated these measures based on a minimum sample size submitted as guidelines by the developer; it was not required. Specifically, the developer noted that measure testing demonstrated reliability with a minimum sample size of 400. NQF endorsement criteria for resource use and quality measures do not require a minimum sample size for measure endorsement. The resource use measure submission process allows developers to submit this information as specifications, guidelines or not at all. The Committee agreed that measure developers need to demonstrate adequate testing and results and considered: (1) if the developers testing was appropriate for the specified measure; (2) if the scope of testing including the representiveness and sample size was adequate; and (3) if the results indicate an acceptable level of reliability and validity. The NQF endorsement criteria are not prescriptive on the type of testing approach or any cut-off for reliability testing scores.

Further, the Committee recognizes that the use of administrative claims data presents certain limitations for measuring resource use performance; these limitations are present in quality performance measurement as well. While administrative data are the primary data source used for measuring resources at this time, the Committee encourages developers to integrate the data gathered through EHRs and other clinical data to measure resource use.

(1609) ETG based hip/knee replacement cost of care measure (Ingenix)
Description: Some commenters expressed support of this measure, noting the measure's ability to capture actual costs at the individual clinician level. Another commenter questioned the measure's clinical logic since this hip fracture measure is based on a non-representative population and the developer submission lacks information on why low-cost outliers are excluded, but high cost outliers were windsorized. Further, the measure fails to capture important and costly complications of comorbidity such as post-operative delirium, pulmonary embolus or dementia.

Committee Response: Concerns related to the clinical logic related to this measure were considered in TAP and Steering Committee discussions; however, the Committee determined that the recommendation for this measure should remain.

(1611) ETG based pneumonia cost of care (Ingenix)
Description: Commenters expressed concern over the validity of the clinical logic, specifically identifying the measure population using administrative claims data with limited ability to distinguish between different types of pneumonia. The inability to distinguish between community-acquired and healthcare-acquired pneumonia will result in the inclusion of costs for episodes of very distinct types of pneumonia into this measure. Further, commenters also believed that there was insufficient information provided to the TAP to determine scientific acceptability. Other commenters disagreed that inclusion of costs six months prior to the pneumonia episode is an inappropriate approach to assigning costs.

Committee Response: The Committee considered the TAP discussion and concern regarding the inability to distinguish between different types of pneumonia. However, ultimately they agreed that this measure should be recommended noting the current limitations of administrative data, limitations that would apply to quality measures as well. The Committee considered concerns on inclusion of six months of costs prior to the pneumonia episodes but determined that the recommendation for this measure should remain.

Measure Specific Comments on the Split Vote Measure

(1595) ETG based diabetes cost of care measure (Ingenix)
Description: Commenters were generally supportive of this measure. One commenter encouraged the Committee and developers to further understand and describe the risk adjustment/stratification approach to ensure that comparisons are reasonable and accurate.

Committee Response: The Committee's initial vote on this measure resulted in a split vote, however, it was agreed that re-voting or reconsidering the measure would likely not result in a substantial difference in Committee stance on the measure. As such, the

Committee determined that the split vote should remain and be forwarded to the membership for vote and the CSAC as is.

Comments on Measures Not Recommended

(1591) ETG-based congestive heart failure (CHF) cost of care measure (Ingenix)
(1594) ETG-based coronary artery disease (CAD) cost of care measure (Ingenix)
(1599) ETG-based non-condition specific cost of care measure (Ingenix)
(1603) ETG-based hip fracture cost of care measure (Ingenix)
(1605) ETG-based asthma cost of care measure (Ingenix)
(1608) ETG-based chronic obstructive pulmonary disease cost of care measure (Ingenix)

Description: Commenters expressed concern over the Committee's decision not to recommend these measures. Commenters believe that all of these measures meet the NQF criteria and should be recommended for endorsement. They also suggest the Committee's rationale for not recommending endorsement for these measures was insufficient.

Committee Response: The Committee considered each measure submitted to this project individually. The Committee encouraged identifying specific supportive or clarifying information related to the clinical logic and construction logic concerns raised. All measures recommended for use as a national consensus standard must meet the same four criteria as quality measures; specifically, importance to measure and report, scientific acceptability of the measure properties, usability and feasibility. Further, the Committee agreed that all measures must meet current standards for reliability and validity testing outlined by the NQF Measure Testing and Evaluation Scientific Acceptability of Measure Properties Task Force report. As such, the Committee determined that the initial recommendation for these measures should remain.

NQF MEMBER VOTING

VOTING RESULTS

Voting results for the four candidate consensus standards are provided below.

Measure #1560 Relative Resource Use for People with Asthma

Measure Council	Yes	No	Abstain	Total Votes	% Approval*
Consumer	4	0	0	4	100%
Health Plan	3	0	0	3	100%
Health Professional	0	0	2	2	
Provider Organizations	0	1	0	1	0%
Public/Community Health Agency	0	0	0	0	
Purchaser	6	1	0	7	86%
QMRI	2	0	0	2	100%
Supplier/Industry	0	0	1	1	

All Councils	**15**	**2**	**3**	**20**	**88%**
Percentage of councils approving (>50%)					80%
Average council percentage approval					77%

*equation: Yes/ (Total - Abstain)

Comments: No comments received

Measure #1561 Relative Resource Use for People with COPD

Measure Council	Yes	No	Abstain	Total Votes	% Approval*
Consumer	4	0	0	4	100%
Health Plan	3	0	0	3	100%
Health Professional	0	0	2	2	
Provider Organizations	0	1	0	1	0%
Public/Community Health Agency	0	0	0	0	
Purchaser	6	1	0	7	86%
QMRI	1	1	0	2	50%
Supplier/Industry	1	0	0	1	100%
All Councils	**15**	**3**	**2**	**20**	**83%**
Percentage of councils approving (>50%)					67%
Average council percentage approval					73%

*equation: Yes/ (Total - Abstain)

Comments: No comments received

Measure #1595 ETG Based Diabetes Cost of Care Measure

Measure Council	Yes	No	Abstain	Total Votes	% Approval*
Consumer	4	0	0	4	100%
Health Plan	3	0	0	3	100%
Health Professional	0	2	0	2	0%
Provider Organizations	1	0	0	1	100%
Public/Community Health Agency	0	0	0	0	
Purchaser	6	1	0	7	86%
QMRI	0	1	1	2	0%
Supplier/Industry	0	1	0	1	0%
All Councils	**14**	**5**	**1**	**20**	**74%**

Percentage of councils approving (>50%)	57%
Average council percentage approval	55%

*equation: Yes/ (Total - Abstain)

Comments:

- ***New Wave:*** Next Wave cannot vote endorsement at this time, but could if our two major concerns are met: Next Wave supports the concepts and structures of the ETG/PEG design to the extent that we were able to review it from the limited data provided to the TAP (the developer Transparency website referenced in the submission only provided marketing level materials - links to reference materials and code maps needed to review substance were broken). The benefit of resource measures is for providers to actually utilize these tools to improve their practices. This depends on the provider understanding the tool sufficiently to trust them. Repairing the broken Transparency links would facilitate this and could be made a condition for CSAC endorsement.
- ***America's Health Insurance Plans (AHIP):***
    1) The diabetes and cardiovascular resource utilization measures were developed by insurance carriers that used complex analyses on large, national insured databases, which mostly excluded adults over age 65 years who are primarily insured through Medicare plans;
    2) More problematic, nursing is virtually absent from all resource use measures, with only individual physician level of analysis and physician practice group and hospital system level of analyses, which may only indirectly represent nursing care/resource utilization using the medical practice model;
    3) As integrated analyses are conducted with the NQF clinical outcomes measures data, it is important to be able to identify outcomes attributable specifically to nursing care and resource use, and this is not possible under the current proposed measures;
    4) All Diabetes/CV TAP measures lack specificity, in that the developers used several overlapping diagnoses that confounded TAP members' ability to determine resource use for individual specific problems associated with and billed for diabetes care. Due much to this lack of specificity, several measures were withdrawn during the Diabetes/CV TAP measurement review meetings, or shortly thereafter by various vendors, including proprietary vendors.

Measure #1609 ETG Based Hip/Knee Replacement Cost of Care Measure

Measure Council	Yes	No	Abstain	Total Votes	% Approval*
Consumer	4	0	0	4	100%
Health Plan	3	0	0	3	100%
Health Professional	0	1	1	2	0%
Provider Organizations	0	0	1	1	
Public/Community Health Agency	0	0	0	0	
Purchaser	6	1	0	7	86%

QMRI	1	1	0	2	50%
Supplier/Industry	0	1	0	1	0%
All Councils	**14**	**4**	**2**	**20**	**78%**
Percentage of councils approving (>50%)					50%
Average council percentage approval					56%

*equation: Yes/ (Total - Abstain)

Comments: No comments received

Measure #1611 ETG Based Pneumonia Cost of Care Measure

Measure Council	Yes	No	Abstain	Total Votes	% Approval*
Consumer	4	0	0	4	100%
Health Plan	3	0	0	3	100%
Health Professional	0	1	1	2	0%
Provider Organizations	1	0	0	1	100%
Public/Community Health Agency	0	0	0	0	
Purchaser	6	1	0	7	86%
QMRI	1	1	0	2	50%
Supplier/Industry	0	1	0	1	0%
All Councils	**15**	**4**	**1**	**20**	**79%**
Percentage of councils approving (>50%)					57%
Average council percentage approval					62%

*equation: Yes/ (Total - Abstain)

Comments: No comments received

General Comment:

America's Health Insurance Plans (AHIP): AHIP supports the endorsement of the Phase II Resource Use Cycle 2 Standards. We would like to provide supplemental comments to our submitted vote. They are as follows:

We believe that resource use measurement and NQF's report could be strengthened in the following ways:

1) We underscore the importance of measures of actual cost of care. It is essential that there be measures of the total cost of care that are based on actual prices paid. Providers who elect to use higher cost services may use the same number of resource units as another provider having similar resource unit expenditures but can result in significant differences in total dollars spent as the unit price of the service may vary. Both actual and standardized cost of care resource use measures are important and preference for resource use versus cost of care measures should not be given;

2) With respect to measures of resource use NQF should explore alternative approaches to get to standardization that do not rely on standardized pricing and would allow comparison across regions. Such alternatives could include use of a standardized comparative denomination such as Relative Value Units (RVUs when available). The value of a Relative Resource Use (RRU) measure would still be comparable across providers, health plans, or other measured populations using a non-monetary denomination (e.g., RVUs). By avoiding a dollar value, the urge to make an inference about cost is removed and appropriate comparisons of only utilization (not cost) can be made;
3) the usability of these measures needs to be further examined by NQF to more accurately reflect the needs of end users such as employers and health plans; and
4) ensure that measures are less subject to proprietary issues and can be more widely used by various healthcare stakeholders.

Regarding measures not recommended for endorsement, measure developers should have an opportunity to address Committee concerns to allow reconsideration.

APPENDIX

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).
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 : 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. Could this measure be improved by including other diagnostic criteria to ensure all appropriate asthma patients are captured?
2. How have you come up with the age strata in your risk-adjustment?
3. Can secondary diagnosis be taken into account within the measurement year?
4. Is cost during the measurement year part of the risk-adjustment strategy?
5. Are your measure results published publically?

Developer Response:

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.
2. The age strata for risk-adjustment are designed around known utilization patterns and clinical treatment patterns.
3. *All* costs for anyone with asthma are counted.
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.

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

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

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.

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.

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)]
Overall Reliability: H-11; M-3; L-0; I-0
Overall Validity: H-4; M-10; L-0; I-0
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:
3a. Measure performance results are publicly reported: H-9; M-0; L-0; I-0
TAP Discussion: The TAP believes this subcriterion has been met as NCQA does extensive audits of their material on a regular basis.
3b. Measure results are meaningful/useful for public reporting and quality improvement: H-5; M-4; L-0; I-0
TAP Discussion: The TAP feels the results are usable and understandable.
3c. Data and results can be decomposed for transparency and understanding: H-6; M-3; L-0; I-0
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.
3d. Harmonized or justification for differences: N/A

Overall Usability: H-7; M-7; L-0; I-0
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:
4a. Data elements routinely generated during care process: H-9; M-0; L-0; I-0
***TAP Discussion*:** The TAP believes this subcriterion has been met as data is a byproduct of care.
4b. Data elements available electronically: H-9; M-0; L-0; I-0
***TAP Discussion*:** The TAP believes this subcriterion has been met; all data is available electronically.
4c. Susceptibility to inaccuracies/ unintended consequences identified: H-6; M-3; L-0; I-0

***TAP Discussion*:** The TAP believes this subcriterion has been met.
4d. Data collection strategy can be implemented: H-8; M-1; L-0; I-0
***TAP Discussion*:** The TAP believes this subcriterion has been met.

Overall Feasibility: H-10; M-4; L-0; I-0
Committee Discussion: There were no new additional comments from the Steering Committee relating to feasibility of NCQA measures.

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.
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.

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.

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.

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.

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.

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.

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.

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.

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).
Resource Use Measure Type: Per episode
Data Source: 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, 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
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:
1a.High Impact: H-9 ; M-0 ; L-0 ; I-0
TAP Discussion: The TAP believes this is a high cost, impact aspect of healthcare; this subcriterion has been met.
1b. Resource use/cost problems: H- 3 ; M-6 ; L-0 ; I-0
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.
1c. Purpose clearly described: H- 4 ; M-5 ; L-0 ; I-0
TAP Discussion: The TAP believes that the intent provided not specific to this diabetes measure, it is a very general statement.
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-18, N-0
Committee Discussion: The Steering Committee believes this is a high impact area that should be

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

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

1595: ETG Based Diabetes Cost of Care Measure (Ingenix)
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 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. **3b. Measure results are meaningful/useful for public reporting and quality improvement: H- 0 ; M-4 ; L-2 ; I-2** ***TAP Discussion:*** The usability information submitted is not specific to diabetes, but for all Ingenix measures. **3c. Data and results can be decomposed for transparency and understanding: H- 1 ; M-2 ; L-5 ; I-0** ***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. **3d. Harmonized or justification for differences: H-0 ; M-0 ; L-0 ; I-0; N-9** ***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 **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: **4a. Data elements routinely generated during care process: H- 8 ; M-0 ; L-0 ; I-0** ***TAP Discussion:*** The TAP agrees this subcriterion has been met; measures rely on administrative data. **4b. Data elements available electronically: H-8 ; M-0 ; L-0; I-0** ***TAP Discussion:*** The TAP agrees this subcriterion has been met; administrative data are in electronic

1595: ETG Based Diabetes Cost of Care Measure (Ingenix)
format. **4c. Susceptibility to inaccuracies/ unintended consequences identified: H-2 ; M-2 ; L-4 ; I-0** ***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. **4d. Data collection strategy can be implemented: H- 5 ; M-2 ; L-1 ; I-0** ***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** ***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)