# COST AND RESOURCE USE MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

**To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return**

## Brief Measure Information

**NQF #:** 1598

**Measure Title:** Total Resource Use Population-based PMPM Index

**Measure Steward:** HealthPartners

**Brief Description of Measure:** The Resource Use Index (RUI) is a risk adjusted measure of the frequency and intensity of services utilized to manage a provider group's patients. Resource use includes all resources associated with treating members including professional, facility inpatient and outpatient, pharmacy, lab, radiology, ancillary and behavioral health services. A Total Cost of Care Index (NQF-endorsed #1604) when viewed together with the Total Resource Use measure provides a more complete picture of population based drivers of health care costs

**Developer Rationale:** By measuring population based relative resource use, health plans and providers can improve the affordability of health care without sacrificing quality. HealthPartners' RUI gives provider groups valuable information on resource use and, when viewed in conjunction with quality metrics, information on the efficiency of care. The HealthPartners RUI measure is a population-based, patient-centered, total resource use measure, created with Total Care Relative Resource Values that cross all categories of health services. This is in contrast to the many, episodic based resource use measures available in the market today. Both population based and episodic based resource use measures are important and complimentary but a key benefit of population based measures is helping to better understand potential overuse & underuse (e.g., although efficient at spine surgery, may be performing too many).

**Resource Use Measure Type:** Per capita (population- or patient-based)

**Data Source**: Claims (Only)

**Level of Analysis:** Clinician : Group/Practice, Population : Community, County or City

**Costing Method:** Standardized pricing

Total Resource Use measure uses the Total Care Relative Resource Values (TCRRVs). TCRRVs are a grand linear scale of relative values designed to evaluate resource use across all types of medical services, procedures and places of service. The values are independent of price and can be used to evaluate providers, hospitals, physicians and health plans against their peers on their efficiency of resource use in treating like conditions.

**Tested Population:** The validity and reliability testing of the measures was conducted with HealthPartners' commercial population which is 470,000 members. For purposes of testing income disparities for the SES analysis, Medicaid was included in addition to commercial which is the combined total membership of 530,000 members.

**Resource Use Service Categories:**

**Attribution Approach**

The level of analysis for this measure could be an entire health plan, provider group, employer group and/or geographic in nature. Measure was tested using commonly used Attribution Algorithm in an open access market (plurality model, using most recent visit as a tie breaker):

• Include twelve months based on first date of service for the measurement year (e.g. January 1 – December 31) of professional

claims experience, with three months of paid claims run out to allow for claims lag.

• Exclude all services that are not office based

• Exclude convenience care clinic visits and hospice services

• Exclude a providers that are not a physician, physician assistant or nurse practitioner

- Assign each service line a specialty based on the servicing physician’s practicing specialty or credential specialty if practicing specialty is not available.
- Include only the following specialties:

- Family Medicine, Internal Medicine, Pediatrics, Geriatrics, OB/GYN

**IF Endorsement Maintenance – Original Endorsement Date:** Jan 31, 2012 **Most Recent Endorsement Date:** Jan 31, 2012

## Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measure still meets the NQF endorsement criteria (“maintenance”). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

### Criteria 1: Importance to Measure and Report

### 1a. High Priority

**1a. High Priority**. This requirement involves demonstrating that the measure focus addresses one of the following:

- A specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF.
- A demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality).

**Summary of information provided to fulfill the High Priority requirement**

- To demonstrate the importance of measure resource use, the developers cite data demonstrating healthcare spending constitutes a high proportion (17%) of the United States gross domestic product (GDP) and high healthcare costs contributes to adults forgoing healthcare.
- The developers suggest that this measure can support a comprehensive measurement system to identify areas of overuse.

**Preliminary rating for High Priority:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

### 1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
### Maintenance measures – increased emphasis on gap and variation

**1b. Performance Gap.** This requirement involves demonstrating a resource use or cost problems exist and there is an opportunity for improvement (i.e., data demonstrating variation in the delivery of care across providers and/or population group (disparities in care)).

- The developer provided performance data from 2015 dates of service from the multi-stakeholder community collaborative, Minnesota Community Measurement (MNCM) that measured the Total Resource Use of 257 provider groups, representing 1.5 million patients receiving care. MNCM found that risk-adjusted medical group resource use had variation up to 55 percent, from 22% below the state average to 33% above the state average. Resource use is presented relative to the state-wide average.
- It is unclear if the performance gap demonstrated is based on the measure as specified.
  - *The developer has clarified that these analysis used the measures as specified.*

**Disparities**

- To examine differences in measure scores by age and gender, the developer examined the distribution of scores in single specialty obstetric and pediatric groups. Data from these analyses were not provided, but the developer states scores were uniformly distributed and not clustered.

***Questions for the Committee:***

- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

**Preliminary rating for opportunity for improvement:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

**1c. Measure Intent**

**1c. Intent of the resource use measure.** This requirement involves describing the measure intent of the resource use measure and the measure construct.

- The intent of this measure is to allow measure implementers to better understand and measure overuse and underuse to drive person-centered management and accountability.
- A population-based measure complements condition and episode-based measures for a complete view of utilization across the measurement year.

***Questions for the Committee:***

- *Is the measure clearly described?*
- *Is it appropriate to measure resource use in this way? At this level of analysis?*
- *Are the costs included appropriate and consistent with the measure intent?*
- *Is there at least one thing that providers can do to achieve a change in the measure results?*

**Preliminary rating for measure intent:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)**

***1a. High Priority***

Comments:
**Measure addresses a topic of clear importance for the US healthcare system. By focusing on the intensity of resource use and removing variation from negotiated prices, this measure is a helpful guide for identifying improvement opportunities.
**Cost is definitely high impact, so the topic meets High Priority definition
**I agree that accurately measuring resource use in a manner that allows accurate and fair comparisons is a high priority.
**Yes, measure meets sub-criterion. NQF assessment captured adequately.
**NOTE: This is a companion measure to 1604 TCI risk adjusted variation in PMPM spending. This measure differs from 1604 in using standardized prices to construct the provider scores and comparison statistics, rather than actual prices. Most of the issues of performance gap, reliability, validity, feasibilty and use are similar for both measures, and where this is the case I have referenced the survey review for 1604, choosing to highlight in this survey differences between the measures.
**Summary resource use measures for primary care are important assessments of crucial domain within the IOM Quality model. Using standardized values applied to utilization per patient provides a nice way to directly compare summarized differences in use rates across providers.
**High

***1b. Performance Gap***

Comments:
**There appears to be wide variation in performance. Because the measure was developed for use within commercially insured populations, there are fewer data for examining disparities. An analysis did demonstrate that clinical risk adjustment had a much larger effect than income variation.
**I am less confident that this metric is providing evidence of a gap in care. The Developers note that the measure is being widely used, but offer only Minnesota and Wisconsin as examples. Shouldn't there be greater data available for a maintenance measure?
**In the example provided under 1b, the measure developers state that the "MNCM found that resource use had variation of up to 55%". Are the measure developers saying that risk adjusted resource use varied by 55% across all medical groups? This seems to be what they are saying, but I would like clearer examples of how this measure demonstrates a gap in care. A measure of this type could help identify excess variation in resource use, thus identifying

a gap in care.  The measure developers do not provide examples of how this measure identifies disparities in care. Applying their measure to a Medicaid population, instead of restricting it to commercial populations, would be more likely to demonstrate disparities in resource use in underserved populations.
**Yes, measure meets sub-criterion. MNCM found that risk-adjusted medical group resource use had variation up to 55 percent, from 22% below the state average to 33% above the state average. Resource use is presented relative to the state-wide average.
Other evidence also shows significant variation and thus room for improvement, e.g. Dartmouth Atlas. Uniform distribution for age and gender for some practice groups.
**The measure shows substantial variation across provider groups, with a substantial variation in prices paid contributing to this variation. No analysis of disparities in care.
**Variability is seen across primary care clinics indicating potential opportunities for improvement.
**High--showed variation

## Criteria 2: Scientific Acceptability of Measure Properties

### 2a. Reliability

### 2a1. Reliability  Specifications
**Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures**

**2a1. Specifications.** This requirement involves providing the full specifications for the measure so that it can be implemented consistently within and across organizations and allow for comparability. Electronic health record (EHR) measure specifications are based on the quality data model (QDM).

Data source(s):

- Claims (measure and risk adjustment model)

**Specifications:**

- This per capita (population- or patient-based) measure calculates total resource use associated with treating members including professional, facility inpatient and outpatient, pharmacy, lab, radiology, ancillary and behavioral health services and is expressed as a ratio.
- To interpret, a score greater than 1.00 indicates higher risk adjusted resource use, compared to a peer group average; a score less than 1.00 indicates less risk adjusted resource use, compared to a peer group average.
- Developer defines peer groups as a group of members, providers, geographic regions or any grouping of member data. The Resource Use measure will return a value that will be relative to the peer group average (e.g., 1.10 = 10% higher than the peer group average).
- The numerator is calculated as the sum of (Total Medical TCRRV / Medical Member Months) + (Total Pharmacy TCRRV / Pharmacy Member Months).The Johns Hopkins Adjusted Clinical Grouper (ACG) risk score is the measure's denominator.

- The developer provides the following steps regarding the measure's construction logic:
  - Obtain all claims that have a date of service in the measurement year. The measurement year is not explicitly defined by the developer, but they provide an example year as running from January 1st to December 31st.
  - Include members enrolled for a minimum of 9 months in the measurement year
  - Include commercial population only
  - Attribution – the developer acknowledges the attribution approach used by measure implementers may vary according the implementer's business purposes and unit of measurement, but does provide the following attribution guidelines:
    - Include twelve months based on first date of service for the measurement year (e.g. January 1 – December 31) of professional
    - claims experience, with three months of paid claims run out to allow for claims lag.
    - Exclude all services that are not office based
    - Exclude convenience care clinic visits and hospice services
    - Exclude a providers that are not a physician, physician assistant or nurse practitioner
    - Assign each service line a specialty based on the servicing physician's practicing specialty or

credential specialty if practicing specialty is not available.

        - Include only the following specialties:
            - Family Medicine, Internal Medicine, Pediatrics, Geriatrics, OB/GYN
    - Costing method – Per the developer, Total Care Relative Resource Values (TCCRVs) TCRRVs are a grand linear scale of relative values designed to evaluate resource use across all types of medical services, procedures and places of service. For this measure, TCRRVs are applied at the procedure level for each component of care with the exception of inpatient, which is applied at the full admission level. The TCRRV weights that are applied to the claim is tested for accuracy and a total TCRRV is calculated.
- Missing data
    - For members that have their pharmacy benefits carved-out, a proxy of the provider's risk-adjusted pharmacy costs is included. This allows for a calculation of total PMPM.
    - For additional carve outs, the developer indicates the "lowest common denominator principle" should be applied, meaning all services carved out of one segment of input data should be carved out of the measure for all segments of input data and all input components (e.g., PMPMs, attribution, and risk adjustment).
- Clinical Logic
    - The developer states clinical logic is not applicable given this is a population-based measure that applies to all care settings and conditions. The developer does not include explicit inclusion criteria in the measure submission.
- Adjustments for comparability: the developer used the following exclusion criteria and risk adjustment approach. The developer does not include explicit inclusion criteria in the measure submission.
    - Exclusion criteria:
        - Members over age 64
        - Members under age 1
        - Member enrollment less than nine months during the one year measurement time window
        - Members who are not attributed to a primary care provider
        - Dollars per member above $125,000
    - Risk Adjustment
        - The measure is risk adjusted for age, gender, and diagnosis using the Adjusted Clinical Group (ACG) method.
        - The ACG method involves:
            - Grouping International Classification Diagnosis (ICD) diagnosis codes into 32 diagnosis groups (i.e., Aggregated Diagnosis Groups (ADGs)). These ADGs are clinically similar and expected to have similar need for healthcare resources.
            - Adjusted Clinical Groups (ACGs) are created from the ADG assignments and are defined by morbidity, age, and sex. Individual members are then assigned to a single ACG category, which quantifies their risk.

**Changes to specifications since previous evaluation:**

- The developer reported one change to the measure specifications. Previously, members were if their total medical and pharmacy costs exceeded $100,000. The developers increased this amount to $125,000 to account for the natural rise in healthcare costs over the past several years.

***Questions for the Committee***:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*
- *Is the clinical logic clear? Is the construction logic clear?*

## 2a2. Reliability Testing Testing attachment
## Maintenance measures – less emphasis if no new testing data provided

**2a2. Reliability testing:** This requirement involves demonstrating that the measure data elements are repeatable, produce the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

**For maintenance measures, summarize the reliability testing from the prior review:**

- In the 2012 submission (Appendix A), the developer provided a summary of the measure score reliability testing conducted using data obtained HealthPartners' primary care Twin Cities metro area providers for the calendar years of 2007, 2008, and 2009. The testing sample included 19 individuals providers and 268, 912 (2007), 272, 491 (2008), and 303, 639 (2009) members. The methods for assessing measure score reliability included – a 90% random sample, a bootstrapping technique, and analysis of performance overtime. In the 90% and bootstrapping methods, reliability was measure as the mean of the variance between the sampling results and the actual results. Results from each method are summarized below:
  - 90% Sample – Variance ranging from -0.00449 to 0.00125 in 2009
  - Bootstrapping – Variance ranging from -0.00473 to 0.00105 in 2009
  - Provider performance - Relatively consistent between 2008 and 2009 with an average difference of 0.0125.
- In the 2012 review, the Committee found the testing adequately demonstrated the measure's reliability and passed the measure on the reliability criterion (High-10; Moderate-6; Low-0; Insufficient-1; NA-0). The Committee did not raise any specific issues with respect to the measure's reliability testing.

**Describe any updates to testing:**

- For this maintenance submission, validity and reliability testing of the measures was conducted with HealthPartners' commercial population which is 470,000 members. (see testing details below).

**SUMMARY OF TESTING**

**Reliability testing level** ☒ **Measure score** ☐ **Data element** ☐ **Both**

**Reliability testing performed with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**

**Method(s) of reliability testing**

Updated testing

- To demonstrate measure score reliability, the developer conducted the following analyses:
  - Comparing actual measure scores to scores calculated by two sampling methods:
    - Bootstrapping
    - A 90% random sample

**Results of reliability testing**

Updated Testing:

- The variances from Actual RUI ranged from -0.0036 to 0.0065 in the bootstrap to -0.0020 to 0.0015 in the 90% sample.

| **Testing Method** | **Results**<br>**Difference between Actual Score & Sampling Scores** | **Results**<br>**Variation** |
|---|---|---|
| Bootstrapping | Range: -0.0036 to 0.0065 | Within groups <1%; Between groups >110% |
| 90% Sample | Range: -0.0020 to 0.0015 | N/A |

***Questions for the Committee:***

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

**Guidance from the Reliability Algorithm Precise specifications (Box 1) → Empiric reliability testing (Box 2) → Score-level testing (Box 4) → Appropriate method (Box 5) → High certainty that measure results are reliable (Box 6a)**

**Preliminary rating for reliability:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

**2b. Validity**

**Maintenance measures – less emphasis if no new testing data provided**

**2b1. Validity: Specifications**

**2b1. Validity Specifications:** This requirement involves demonstrating that the measure specifications are consistent with the measure intent described under criterion 1c and capture the most inclusive target population.

**Specifications consistent with intent described in 1c.** ☒ **Yes** ☐ **Somewhat** ☐ **No**

***Question for the Committee:***

- o *Does the Committee agree the specifications are consistent with the intent of the measure?*
- o *Is the attribution approach consistent with the measure intent?*
- o *Does the accountable entity have reasonable control over the resources measured?*

**2b2. Validity testing**

**2b2. Validity Testing** This requirement involves demonstrating that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided.

**For maintenance measures, summarize the validity testing from the prior review:**

- In the 2012 submission, the developer provided a summary of construct validity testing conducted using data obtained HealthPartners' primary care Twin Cities metro area providers for the calendar years of 2007, 2008, and 2009. The testing sample included 19 individuals providers over 300,000 members in the 2009 sample. Construct validity was tested by examining the correlations between the measure score and known utilization metrics and ACG scores.
- In the 2012 review, the Committee passed the measure of validity testing (High-5; Moderate-8; Low-2; Insufficient-1; NA-0). The Committee clarified that because this measure has only been tested in a commercial population, it will be NQF endorsed only in a commercial population.

**Describe any updates to validity testing:**

- For this maintenance submission, the developer summarized updated validity testing conducted using provider data from 2014 and 2015 The validity and reliability testing of the measures was conducted with HealthPartners' commercial population which is 470,000 members. This updated validity testing consisted of correlations the measure components (i.e., ACG scores, unadjusted costs) and measure score with other markers of utilization.

**SUMMARY OF TESTING**

**Validity testing level** ☐ **Measure score** ☐ **Data element testing against a gold standard** ☒ **Both**

**Method of validity testing of the measure score:**

☐ **Face validity only**

☒ **Empirical validity testing of the measure score**

**Validity testing method:**

- *Data element validity*
  - o demonstrate data element validity, the developer conducted a series of correlation analyses:
    - ▪ Measure components (i.e., ACG scores & Non-Risk Adjusted Total Cost Relative Resource Values (TCRRVs))
      - ACG Risk-adjusted Total Cost Index (i.e., the measure score)
      - ACG risk-adjusted Resource Use Index (RUI) (i.e., measure 1598)
      - Non-risk adjusted Total Cost Relative Resource Values (TCRRVs)
      - Price
    - ▪ Measure component - Non-Risk Adjusted TCRRVs with non-risk adjusted rates of utilization:
      - Inpatient Admits per 1,000
      - ER per 1,000
      - Outpatient surgery per 1,000
      - High Tech Radiology per 1,000
      - E&Ms per 1,000
      - Lab/Path per 1,000
      - Standard radiology per 1,000

            - Pharmacy per 1,000
        - Measure Components with Composite Utilization
- *Measure score validity – Empirical Testing*
    - To demonstrate measure score validity, the developer conducted a series of correlation analyses:
        - ACG Risk-adjusted Risk Use Index (i.e., the measure score) with:
            - Hospital based Total Cost of Care Index
            - Professional Total Cost of Care Index
            - Pharmacy Total Cost of Care Index
            - ACG risk-adjusted Total Cost Index (i.e., measure 1604)
            - Total Price
        - Service Category RUI (i.e., Inpatient, Outpatient, Professional, Pharmacy) with risk-adjusted service category metrics:
            - Inpatient admit rate
            - ER count
            - Outpatient surgery
            - High tech Radiation
            - E&M Visits
            - Lab/Path
            - Standard Radiology
            - Prescription (Rx) Count
        - Measure Score with Composite Utilization
        - Measure Score Over Time
- *Measure score validity – Face Validity*
    - To demonstrate measure score face validity, the developer cites their process of sharing measure scores and measure methodology with measured providers.
    - NQF requires a systematic assessment of face validity to be assessed. A systematic assessment of face validity is used when a panel of experts evaluates the measure specifications and measure testing to assess if the measure is an accurate reflection of performance. Results from a panel of experts is not included.
    - *Additional face validity information provided by the developer:*
        - *HealthPartners measures have been systematically evaluated for face validity by the following organizations, each convening panels of experts:*
        - *HealthPartners:  Internally reviewed by Cost Assessment Committee (medical directors, network management, health informatics).  Since 2010, transparent quarterly reporting to 60+ provider groups in the HealthPartners network.  All providers have 45 days to review prior to public reporting.*
        - *Minnesota Community Measurement:  Reviewed by two multi-stakeholder groups - Cost Technical Advisory Group (TAG) – including patients, providers and purchasers, and the Measurement and Reporting Committee (MARC)  - consumers, providers, health plans, purchasers prior to public reporting*
        - *Total Cost of Care – measure used as specified (pages 1-5):  http://mncm.org/wp-content/uploads/2013/04/2014.11.12-MARC-Minutes_Approved.pdf*
        - *Total Resource Use – measure used as specified, known as 'RRU' in this document, (pages 2-3): http://mncm.org/wp-content/uploads/2016/11/2016.09.14-MARC-Minutes_Approved.pdf*
        - *Network for Regional Healthcare Improvement (NRHI) – The RWJF grant funded to produce and distribute practice level regional Total Cost of Care Reports.  The first phase represented five regional health care improvement collaboratives in Colorado, Maine, Missouri, Minnesota and Oregon.  Each region produced and distributed practice level reports in their communities, and a benchmark approach was developed and tested, and have committees and board of directors that oversee the work.*

**Validity testing results:** (highlighted value are those directly relevant to the measure under evaluation)

- *Data element validity testing results*
    - Correlation between measure components, ACG Score and Non-Risk Adj PMPMs and other metrics

| | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **ACG** | **Non-Risk Adj PMPMs** |
| Non-Risk Adj PMPM | 0.62 | 1.00 |
| Non-Risk Adj TCRRVs | 0.88 | 0.78 |
| ACG Risk Adj TCI | 0.03 | 0.79 |
| ACG Risk Adj RUI | 0.14 | 0.45 |
| Price | -0.09 | 0.57 |

    - The developer notes that there is high correlation of the measure components to one another and each component’s correlation with the Non-Risk Adj TCRRVs as sufficient evidence for the validity of the measure components.
        - The correlation between the non-risk adjusted PMPM and the ACG risk adjusted RUI is 0.45.
        - The developer attributes the low correlated between ACG and Price to fact that ACG is an estimate of expected resource use whereas price is the unit cost of services actually provided.

    - Measure component - Non-Risk Adj PMPMs with non-risk adjusted rates of utilization:

| **Non-Risk Adjusted** | **Correlation Coefficient** | |
|---|---|---|
| ***Service Category*** **Metric** | ***Non-Risk Adj Service Category PMPMs*** | ***Non-Risk Adj Service Category TCRRVs*** |
| ***Inpatient*** | | |
| Admits/1000 | 0.67 | 0.82 |
| ***Outpatient*** | | |
| ER/1000 | 0.67 | 0.52 |
| OP Surgery/1000 | 0.60 | 0.68 |
| HighTech Rad/1000 | 0.45 | 0.67 |
| ***Professional*** | | |
| E&M/1000 | 0.63 | 0.71 |
| Lab/Path/1000 | 0.77 | 0.83 |
| Std Rad/1000 | 0.49 | 0.72 |
| ***Pharmacy*** | | |
| Rx/1000 | 0.73 | 0.80 |

    - Measure Components with Composite Utilization

| **Non-Risk Adjusted** | **Correlation Coefficient** | | |
|---|---|---|---|
| **Metric** | **ACG** | **Non-Risk Adj PMPMs** | **Non-Risk Adj TCRRVs** |
| Composite Utilization | 0.74 | 0.69 | 0.87 |

- *Measure score validity – Empirical Testing*

| Risk Adjusted | Correlation Coefficient | | |
|---|---|---|---|
| Metric | TCI | RUI | Price |
| Hospital TCI | 0.74 | | |
| Prof TCI | 0.73 | | |
| Rx TCI | 0.16 | | |
| Hospital RUI | | 0.30 | |
| Prof RUI | | 0.74 | |
| Total RUI | 0.39 | | |
| Hospital Price | | | 0.86 |
| Prof Price | | | 0.83 |
| Total Price | 0.87 | | |

- Service Category TCI (i.e., Inpatient, Outpatient, Professional, Pharmacy) with risk-adjusted service category metrics:

| Risk Adjusted | Correlation Coefficient | |
|---|---|---|
| *Service Category* Metric | *Service Category TCIs* | *Service Category RUIs* |
| ***Inpatient*** | | |
| Admit Rate | 0.78 | 0.82 |
| ***Outpatient*** | | |
| ER Cnt | 0.68 | 0.46 |
| OP Surgery | 0.55 | 0.49 |
| High Tech Rad | 0.21 | 0.37 |
| ***Professional*** | | |
| E&M Visits | 0.48 | 0.70 |
| Lab/Path | 0.59 | 0.54 |
| Std Rad | 0.48 | 0.38 |
| ***Pharmacy*** | | |
| Rx Count | 0.25 | |

- Measure Score with Composite Utilization

| Risk Adjusted | Correlation Coefficient | Correlation Coefficient |
|---|---|---|
| Metric | TCI | RUI |
| Composite Utilization | 0.72 | 0.52 |

- Measure Scores Over Time

| | TCI | | | | Price | | | | RUI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Provider Group Size | 25th Percentile | Average | Median | 75th Percentile | 25th Percentile | Average | Median | 75th Percentile | 25th Percentile | Average | Median | 75th Percentile |
| <1,000 | 0.04 | 0.07 | 0.07 | 0.11 | 0.02 | 0.04 | 0.03 | 0.05 | 0.03 | 0.05 | 0.05 | 0.09 |
| 1,000-2,000 | 0.03 | 0.08 | 0.07 | 0.11 | 0.02 | 0.03 | 0.03 | 0.04 | 0.03 | 0.06 | 0.07 | 0.09 |
| 2,000+ | 0.01 | 0.03 | 0.03 | 0.04 | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 | 0.03 | 0.03 | 0.05 |

***Questions for the Committee:***

- *Is the test sample adequate to generalize for widespread implementation?*

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *For data element validity and measure score validity, are the correlations in the expected direction and of the expected magnitude?*
- *Are the correlations between the measure score and place of service metrics sufficient for demonstrating measure score validity?*

### 2b3-2b7. Threats to Validity

**2b3. Exclusions**: This requirement involves demonstrating that the exclusions are:

- supported by the measure intent

AND/OR

- There is a rationale or analysis demonstrating that the measure results are sufficiently distorted due to the magnitude and/or frequency of the non-clinical exclusions;

AND

- Measure specifications for scoring include computing exclusions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion);

AND

- Patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

**Summarize approach and analysis of exclusions**

- Members with the following characteristics are excluded from the measure:
  - Members over age 64
  - Members under age1
  - Member enrollment less than 9 months during the one year measurement time window
  - Members not attributed to a primary care provider
  - Dollars per member above $125,000 are excluded (i.e., truncated)
    - In the 2012 submission, the exclusion amount was $100,000
- The developers states testing shows the exclusion of members under 1 and those without 9 months of enrollment during the measurement yet has little impact on the model's $R^2$ value, but do not provide specific data to support this.
- Analyses were not conducted examining the effect of excluding Members over 64, rather the developer state they are excluded due to potential incomplete claims data from Medicare eligible beneficiaries.

***Questions for the Committee:***

- *Are the exclusions consistent with the intent of the measure? Are carve-outs appropriately addressed?*
- *Are any patients or patient groups inappropriately excluded from the measure? Specific patient groups to consider include patients who died during the measurement period, patients who were transferred, and patients enrolled in Medicare Advantage plans.*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: This requirement involves specifying an evidence-based risk-adjustment strategy (e.g., risk models, risk-stratification) that is based on patient clinical factors that influence the measured outcome and are present at the start of care and has demonstrated adequate discrimination and calibration. If a risk adjustment strategy is not provided, a rationale or data to support no risk-adjustment/-stratification must be provided.

**Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

**Conceptual rationale for SDS factors included ?** ☒ **Yes** ☐ **No**

**SDS factors other than age and gender included in risk model?** ☐ **Yes** ☒ **No**

**Risk adjustment summary**

- The risk adjustment approach utilized in the measure is the Johns Hopkins Adjusted Clinical Grouper (ACG) method, which adjusts for age, gender, and diagnosis (i.e., clinical risk). A conceptual rationale for this risk adjustment approach is provided.
- The risk adjustment approach involves:
  - Grouping International Classification Diagnosis (ICD) diagnosis codes into 32 diagnosis groups (i.e., Aggregated Diagnosis Groups (ADGs)). These ADGs are clinically similar and expected to have similar need for healthcare resources.
  - Adjusted Clinical Groups (ACGs) are created from the ADG assignments and are defined by morbidity, age, and sex. Individual members are then assigned to a single ACG category, which quantifies their risk.
- Individual member ACG weights: Individuals are assigned to an ACG actuarial cell that has a corresponding weight reflecting relative illness burden. The ACG weight is then multiple by their number of eligible member months.
- Providers' ACG Scores are calculated as the sum of their attributed members ACG weights.
- The developer does not provide a summary of statistical results of the analyses conducted on ACG risk model as that information is proprietary.

**Empirical Summary of SDS**

- Two measures of income - tract-level income, obtained from U.S. Census Tract data, and the household-level, obtained from a commercially licensed consumer database purchased by HealthPartners – were used to examine the impact of SDS on the measure scores.
- Two multiple linear regression equations were analyzed:
  - Equation 1: Tract-level income, ACG risk score, and insurance product (i.e., Commercial vs Medicaid) were regressed on total reimbursed amount per member per month; and
  - Equation 2: Household-level income, ACG risk score, and insurance product (i.e., Commercial vs Medicaid) were regressed on total reimbursed amount per member per month
- Results from both Census tract-level and household-level data sources show that income does not significantly impact the measure scores after risk adjusting for age, gender, and clinical risk, and stratifying by insurance type. The ACG score and the insurance type have a significant impact on the cost and resource use measures' variation and income has no discernible impact.

**Risk Model Discrimination and Calibration**

- For model discrimination, the developers provider the correlations of non-risk adjusted PMPM and ACG scores with other metrics of utilization. Discrimination and calibration statistics were not provided.

***Questions for the Committee:***

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors in their risk-adjustment model?*

2b5. Meaningful difference: This requirement involves demonstrating, through data analysis, that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically meaningful differences in performance.

- To demonstrate the measure's ability to identify meaningful differences, the developer provides additional guidance on interpreting measures scores and a summary of performance in 66 providers groups.
- The developer noted that the methods for scoring does not allow for identification of statistically significant and practically meaningful differences of performance as it is based on a full population.

***Question for the Committee:***

- *Does this measure identify meaningful differences about cost or resource use?*

2b6. Comparability of data sources/methods: This requirement involves demonstrating that if multiple data sources/methods are specified, they produce comparable results.
N/A

2b7. Missing Data: This requirement involves describing how missing data are handled and demonstrating that the presence of missing data does not bias the measure.

- The developer states that this is a full population-based measure and all data is included.
- For members that have their pharmacy benefits carved-out, a proxy of the provider's risk-adjusted pharmacy costs is included. This allows for a calculation of total PMPM .
- For additional carve outs, the developer indicates the "lowest common denominator principle" should be applied, meaning all services carved out of one segment of input data should be carved out of the measure for all segments of input data and all input components (e.g., PMPMs, attribution, and risk adjustment).

**Guidance from the Validity Algorithm** Precise specifications (Box 1) YES → Empirical testing conducted with measure as specified (Box 2) YES → Measure Score validity testing conducted (Box 6) YES → Testing method described and deemed appropriate (Box 7) YES → Moderate certainty that the measure score is a valid indicator of quality

**Preliminary rating for validity:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b)**

***2a1. & 2b1. Specifications***

***2a1. Reliability Specifications***

Comments:
**Reliability is high; the measure is clearer defined and implemented.
**Claims-based measure with widely used commercial risk adjustment tool; specifications appear clear and reliable.
**For sample size, they note that the testing was done with an attributed population of 600 members. Is this within a group? By a payer? By a plan? Not clear. Attribution was a big concern in 2012 when this was last reviewed. I am not sure that concern has been answered.
**Why as a 9 month minimum enrollment period chosen? I would like to see justification for this time period vs. longer or shorter time periods. I would like more information about how the TCRRV weights are calculated.
How do the measure developers calculate the proxy for members whose pharmacy benefits are carved out?
Why was this measure only tested on commercial populations? I think not testing the measure on Medicare and Medicaid populations is very problematic. Almost half of my patient panel is 65+. For medical groups like mine that have large numbers of managed Medicare and Medicare ACO patients, this measure would not be useful.
Which EMR vendors can implement this measure at this time? Would users have to purchase 3rd party products to utilize this measure?
**Yes, measure meets sub-criterion. NQF assessment captured adequately.
**Is Attribution method part of the Total Resource Use Population-based PMPM Index process? If attribution method differs from that proposed, what impact does that have on NQF approval?
In Minnesota Community Measurement application of TCOC measures, problems with their attribution method result in the attribution of patients seen in specialty clinics (e.g., cancer center, GI clinic) by NP/PAs and medical residents prior to their board certifications as primary care patients because these groups are counted as internal medicine providers.
**High--test retest looks good/consistent; no information on discrimination (signal to noise)

***2a2. Reliability Testing***

Comments:
**Okay
**No serious concerns about reliability
**The reliability testing seems solid. I have some worries about risk adjustment that will be detailed below.
**Testing seems adequate for the populations measured.
**Yes. Additional reliability testing presented since last endorsement. Conducted at measure score level.
**Although the population was of good size, it appears that all claims were from one payer and a restricted area of the west north central region of the the US. This part of the country has a larger portion of large group practices than much of the US. Generalizability would be strengthened by including more payers and a wider range of provider types.
**Low/moderate. No results for signal/noise (discrimination). Is 600 enough? What is the reliability to differentiate performance of different providers?

***2b1. Validity Specifications***

Comments:

**Specifications are clear and consistent with evidence.  While there may be validity concerns, they are not in specifications.
**Specifications seem reasonable.  The score is easy to interpret.
**The lack of testing in patients age 65+ limits the measure's validity to patients under age 65. This is a significant drawback. It is unclear from the information provided how medical groups would use the data this measure provides to change their resource use.  First, a broad index only indicates that overuse exists, but does not tell the user where. Second, measure users may have limited control over some of the areas of overuse (ex. hospital care).
**There does not appear to be an inconsistency.
**Use of standardized weights to combine different types of utilization provides a reasonable way to compare providers. How is the greater presence of payment bundling  than when this measure was developed affecting the relative weights that drive this measure? It is not clear how high cost patients are identified or Windsorized.
**Moderate to high---no information on sharing results with providers (i.e., face validity).  Correlates moderately to highish with other measures of use.

***2b2. Validity Testing***

Comments:

**Since it is not clear what the measure is assessing vis a vis resource use, the approach to validity testing is not clear. Overall correlations of the measure with other measures of use or cost demonstrate a weak basis for judging whether the variation in spending is actionable or not, i.e., whether the PMPM cost differences can or should be narrowed.
The risk adjuster is a well-established one, but I would have liked to see more discussion of how much of the variance it explains and the extent to which the rankings change when risk adjustment is introduced.
I would also like to have seen more analysis of variance, breaking allocating the variance in RA PMPM expenses to price differences across groups, and high or low use of specific services.
The high correlation of PMPM and prices suggest that price variations account for a substantial portion of this measure's variance.  How should a payer interpret this? A group?
**Some of the correlation coefficients in the validity testing were modest.   Overall the index appears to be valid; however, the committee will likely want to walk through this in some additional detail.
**Was the measure tested both within and between different specialties?  It is perfectly appropriate to consider OB/GYN as a primary care specialty, but the costs of a proceduralist will likely be much higher than an office-based physician.  Were there differences found between specialties defined as primary care?
**I cannot draw a conclusion as to whether this measure is an indicator of quality as no data on clinical quality is provided.  Only cost data is included. Many of the correlations coefficients quoted fall in the moderate range. I would like to see higher correlation coefficients.
**Yes. Uses validity testing at both measure score and data element. Uses empirical validity testing. This measure can be viewed with TCC and quality to get general understanding of value.
**Except for the previously mentioned issues, validity testing reflects that the TCOC total resource use measure is reflecting the average expenses per patient acceptably well.
**Moderate to high--done in commercial population.  Large # of entities (medical groups of varying sizes)

***2b3. Exclusions Analysis***

Comments:

**The population exclusions look reasonable.  Risk adjustment is done with a widely adopted metric.I would like some discussion of the proportion of groups with drug carveouts and the variance in drug spending among those groups for whom the data are available.  Would also like to know about other carve outs, particularly mental health services, and the proportion of costs these represent. Most important issues involve carve outs.  Their magnitude is not discussed in the documentation.
**Age: <1 ; >64; < 9 months enrollment
Truncation @ $125,000 spend
As specified this accounts for just under 80% of members and total spend.   Does the committee feel this is adequate?
**Exclusions seem reasonable
**I am still bothered that patients over age 64 are excluded.
**This measure excludes (truncates) member medical and pharmacy costs that are over $125,000. The AAMC requests an explanation and rationale as to why these medical and pharmacy costs are capped and why $125,000 was selected as the threshold.  The AAMC also requests an explanation as to why non-provider administered drugs (those not covered under Medicare Part B) are not included in the cost calculation for these measures.
**High cost patient are not excluded but are Windsorized (truncated at the threshold value, moving from $100,000 per

year to $125,000 per year). This threshold appears to be over 20 times the average cost per person per year. This would still result in potential undue influence on the TCRRV measure by a small number of patients. I would recommend that outliers be excluded rather than "capped". The MSPB measure excludes inter-institution transfers because neither institution has full influence on major portions of the costs incurred. Most cost outliers for TCRRV would have the same issues in that much of the cost would be outside the primary care providers control (inpatient or specialty care ordered by others).
**High--no problems

***2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures***
Comments:
**Risk adjustment uses a standard widely adopted measure. The use of census tract level income and other variables is commended. The analysis shows low variance due to SDS variables but this may be due to low variance across the groups included in the analysis of the SDS variables. Would like to see the distribution across groups (not population as a whole) of these measures and better understand the extent of the variance in SDS measures at the group level.
**The committee will want to walk through the SES testing that was performed.
**The risk adjustment presented used the ACG system. Is that publicly accessible? I think it is a proprietary tool that has to be licensed to the groups using it. I am concerned about using an opaque means of risk adjustment in each of these measures. What is used in this measure? I tried to look this up on the provided website reference, but it just wanted to license the product to me. Did the developers look at the differences in risk adjustment values between institutions?
**I would like more detail on the Adjusted Clinical Grouper (ACG) risk adjustment method.
I am concerned that SDS is excluded. I am not sure the rationale the developers provide justifies this exclusion.
**Yes, SDS risk-adjustment variables present at start of care. On average there was less than a 1% change in performance for provider groups when income was introduced into the model for the Resource Use measure when using Census tract data. This impact was reduced on average to less than a 0.25% when using the commercially licensed data source with more specific income data. Yes, well developed risk-adjustment model. The measure uses an ACG-Johns Hopkins risk adjustment methodology which is proprietary. Analysis of risk-adjustment methods and found little difference in results.
**ACG risk adjustment seems like a reasonable approach for TCRRV. It is important to point out that TCRRV is developed for a commercially insured population. In that setting, income has little influence on the measure. When adding Medicaid patients to the analysis, the reimbursement differences between Medicaid and commercial groups is likely confounded with SDS differences.
**Moderate---income variable had modest effect on R2. Do we know whether inclusion changed any provider's rank position in the distribution and by how much? May only affect a small number of providers, but potentially in important ways.

***2b5. Identification of Statistically Significant & Meaningful Differences In Performance***
Comments:
**Recognize the language is standardized, but this is not a quality measure.
There appear to be substantial variation in the PMPM costs, with a substantial portion of this due to differences in the prices the groups get. Would like some committee discussion of the magnitude of the differences across groups.
Also, the range of variation here seems smaller than for the price based 1604, but as with that measure, there is no formal standard for how big a difference. With these complementary measures, one of the limitations is the information on price and volume are not presented in an interesting way. For example, do providers who have low use of hospitalization or ancillary services command higher prices for E&M type services? Can't be answered from the way the data are presented.
**I have difficulty drawing conclusions about quality of care from a resource use measure without a better understanding of how "quality" is defined. How much extra resource use to provide the same quality of care is acceptable and how much is excessive?
**Yes, this measure identifies meaningful differences about resource use. It is based on deviation from average, so one thing to consider is the average changing over time (and does that seem to be an appropriate change).
**The proponents state that they feel no need to identify statistically significant differences since they're looking at populations. However, it would still be useful to patients and providers if they recommended a measure of spread (standard error or standard deviation) to be included along with their point estimates when reporting provider performance.
**Identifies meaningful differences in the use of services

***2b6. Comparability of Performance Scores When More Than One Set of Specifications***
Comments:

**N/A

***2b7. Missing Data Analysis and Minimizing Bias***

Comments:
**Does not appear to be a problem.
**No
**No
**I need a clearer explanation of the pharmacy cost proxy the developers use.
**It is not clear how extensive pharmacy carve-outs are encountered. It is also not apparent how outlier thresholds are adjusted for missing pharmacy data. Was any testing done to assess the reliability and validity of imputing pharmacy data when that component has a "carve-out" benefit?
**High--no problems

**Criterion 3. Feasibility**
**Maintenance measures – no change in emphasis – implementation issues may be more prominent**

**3. Feasibility:** This requirement involves demonstrating:

- the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
- the required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.
- the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

**Data Specifications and Elements**

- The measure is constructed using administrative health claims, which are routinely created and do not create undue burden for measure implementers
- All data elements are available in defined fields within electronic sources
- The measure uses an ACG-Johns Hopkins risk adjustment methodology which is proprietary.

**Data Collection Strategy**

- Data collection strategy can be implemented as it's currently in operational use by HealthPartners

***Questions for the Committee:***

o *Are the required data elements routinely generated and used during care delivery?*
o *Are the required data elements available in electronic form EHR or other electronic sources?*
o *Can the measure be consistently implemented using a proprietary risk adjustment methodology?*

**Preliminary rating for feasibility:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 3: Feasibility**

***3.Feasibility***

Comments:
**Claims based measure. Feasible to implement.
**The measure's use has become more widespread since the original submission. The measure appears to be very feasible.
**The measure uses easily available metrics
**The data elements needed to use this measure reside in different locations and are controlled by different parties. Outside of a totally integrated and closed health system, obtaining complete data on resource use is challenging. The payors with whom we contract provide different amounts of the needed resource use data.
**Uses claims data, so generally very feasible to implement.
**Data elements are routinely collected. More issues exist in how the attribution is applied.
**High

**Criterion 4: Usability and Use**
**Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences**

**4. Usability and Use**: This requirement involves describing the extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

**Current uses of the measure** [from OPUS]

**Publicly reported?** ☒ **Yes** ☐ **No**

**Current use in an accountability program?** ☒ **Yes** ☐ **No** ☐ **UNCLEAR**
**OR**
**Planned use in an accountability program?** ☐ **Yes** ☐ **No**

- The developer states that there are multiple accountability programs and sub-programs that this measure utilizes including:
  - 3 Public reporting programs
  - 1 Payment program
  - 1 Public Health/Disease Surveillance program
  - 5 Quality Improvement with Benchmarking programs (external benchmarking to organizations)
  - Several Quality Improvement with Benchmarking (internal to the specific organization) programs
- The developer also cited measure page views at the National Quality Measures Clearinghouse (NQMC) from Agency for Healthcare Research and Quality (AHRQ)
  - Reported the following usage between 3/1/15 – 2/29/16
    - 5,815 page views for the Total Cost of Care Measure
    - 1,493 page views for the Total Resource Measure

**Improvement results**

- Large number of those who have adopted the measure and resulted in improvement through greater transparency, which allows users to pinpoint areas for improvement and define strategies to reduce those costs
- One specific example is the Northwest Metro Alliance, which serves more than 300,000 people receiving care at 9 different clinics and one hospital, demonstrated that their medical cost increases were more than 31% lower than the Twin Cities metro average for Commercial patients since they adopted the developer's measure in 2010.

**Unexpected findings (positive or negative) during implementation**

- The developer did not note any unexpected findings during the implementation of the measure

**Potential harms**

- The developer is unaware of negative unintended consequences from other organizations utilizing the measure

**Vetting of the measure by those being measured**

- Since endorsement the measure developer have received some general input regarding implementation of the measure. HealthPartner's organized a public-facing website with resources for external organizations on how to download the necessary tools to run the measure.

**Measure can be deconstructed to facilitate transparency and understanding** ☒ **Yes** ☐ **No**

**Feedback:**

***Questions for the Committee***:

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

o *How has the measure been vetted in real-world settings by those being measure or others?*

**Preliminary rating for usability and use:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

### Criteria 4: Usability and Use

***4.Usability and Use***

Comments:

**The developers note that this measure should be used in conjunction with the RCU measure, but the addition this measure offers to understanding of resource use variations is the addition of price. A more useful measure would identify the marginal contribution of price to the PCU measure in explaining variations in resource use, and this is not how the measures are presented. From a user actionability orientation, the reports on these measures provide information by type of service on whether the provider is higher or lower, and this rather than the overall score makes the measure actionable and usable.

Particularly the inclusion of relative rankings on individual service components of the measure. Without these, usability is much lower. I would also stress the issues raised in 2b5 on how the two complementary measures could contribute to understanding of pricing and overall resource use. With these complementary measures, one of the limitations is the information on price and volume are not presented in an interesting way. For example, do providers who have low use of hospitalization or ancillary services command higher prices for E&M type services? Can't be answered from the way the data are presented. Is there other evidence that higher prices are paid when resource use is low or quality is high? Again, can't be answered, it is not clear that the reporting of these measures allow price and resource use measured by standardized prices to be disentangled.

**For geographic analyses the measure is very straightforward. When comparing across provider groups, attribution models come into play and there is not a lot of discussion about how best to attribute members across groups. Obviously this is a topic that creates controversy.

**I would like more information here, again especially considering that this is a maintenance measure. The Developer notes that a number of groups that are collecting data using the measure, but what actual performance data are they seeing? What gaps are being identified? What progress is being made?

**This resource use index seems analogous to body temperature. A fever tells me there is a problem, but not where.

**High for payers and purchasers.

**The measure is used in a variety of programs.

- 3 Public reporting programs
- 1 Payment program
- 1 Public Health/Disease Surveillance program
- 5 Quality Improvement with Benchmarking programs (external benchmarking to organizations)
- Several Quality Improvement with Benchmarking (internal to the specific organization) programs

**Health Partners reports using the measure in multiple ways. It has been used by Minnesota Community Measurement and is reported to be used in other applications across the US.

### Criterion 5: Related and Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

**Related or competing measures**

The developer did not identify and related or competing measures. **5.a. Harmonization**: This requirement involves demonstrating that the measure specifications are harmonized with related measures OR the differences in specifications are justified.

N/A

| Endorsement + Designation<br>The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria. |
|---|
| **This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it:** is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.<br>**Eligible for Endorsement + designation**: ☒ **Yes** ☐ **No**<br>**RATIONALE IF NOT ELIGIBLE**: |

## Pre-meeting public and member comments

- Sia Lo on Behalf of Beth Averbeck from HealthPartners Medical Group on 2/2/17:

HealthPartners Medical Group strongly recommends for endorsement both the Total Cost of Care (#1604) and Total Resource Use (#1598) measures. For more than a decade, Total Cost of Care (TCOC) has been the top line measure of affordability for our care group. We drill down from the overall measure of TCOC to price drivers, and Total Resource Use drivers to identify opportunities for improvement. These measures have guided our improvement strategies; allowing us to focus on appropriate use of services and place of service opportunities. This has resulted in improved affordability for our patients. Our full statement of support and usability of these measures was included in the measure submission.
Nance McClure, Chief Operating Officer and Brian Rank, MD, Executive Medical Director, and Beth Averbeck, MD, Senior Medical Director Primary Care

- Sia Lo on Behalf of Nance McClure from HealthPartners Medical Group on 2/2/17:

HealthPartners Medical Group strongly recommends for endorsement both the Total Cost of Care (#1604) and Total Resource Use (#1598) measures. For more than a decade, Total Cost of Care (TCOC) has been the top line measure of affordability for our care group. We drill down from the overall measure of TCOC to price drivers, and Total Resource Use drivers to identify opportunities for improvement. These measures have guided our improvement strategies; allowing us to focus on appropriate use of services and place of service opportunities. This has resulted in improved affordability for our patients. Our full statement of support and usability of these measures was included in the measure submission.
Nance McClure, Chief Operating Officer and Brian Rank, MD, Executive Medical Director, and Beth Averbeck, MD, Senior Medical Director Primary Care

- Ms. Ellen Gagnon from Network for Regional Healthcare Improvement on 2/21/17:

On behalf of NRHI, we are in support of NQF endorsing this measure. For over three years we have been actively engaged with regions across the country measuring, reporting and using the total resource use population based PMPM index. Recently we published a benchmark report that utilized this measure and compared across 5 regions which has resulted in meaningful conversations within regions about the cause of variation. Seven regions have produced and distributed attributed practice level reports in their communities at least once, some multiple times over the past few years. During 2015, healthcare cost information on over 5 million patients attributed to 20,000 individual physicians were included in practice level reports and used by practices to identify areas of variation and opportunities for intervention to improve care while decreasing costs. The utility of this measure increases as you are able to isolate resource use - which is very powerful and something physicians can control.

The basic foundation for all of these efforts is the HealthPartners NQF endorsed TCOC measure framework. NRHI has been awarded funding from RWJF for a third phase which began on November 1, 2016. During this two-year grant, we will expand the number of regions producing, sharing and using TCOC for both commercial and Medicare populations, maintain and grow our Getting to Affordability Learning Modules and community - a place to connect with others across the country who are measuring and using TCOC, convene a multi-stakeholder summit on using

TCOC to advance the Triple Aim and payment reform, and develop and implement sustainability plans to ensure future ability to produce, share and use TCOC.
We support further endorsement of this measure and would be happy to answer any questions.

- Benson Shih-Han Hsu, MD, MBA, FAAP from Sanford Health on 2/23/17:

Sanford Health supports endorsement of the HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. As an integrated health system in the HealthPartners network, we appreciate the transparency and soundness of the measures, as well as our partnership with HealthPartners as we strive to improve care for our patients. The Sanford Health Plan is also a licensee and user of the measures.

- Steven Mark Connelly, MD from Park Nicollet Health Services on 2/24/17:

Park Nicollet appreciates the opportunity to voice our support for HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. HealthPartners has transparently shared the measurement method and measure results with providers in our community for nearly a decade, and we have used these measures to improve health care affordability for our patients, while maintaining top quality performance. Our full statement of support and comment on the usability and usefulness of these measures was submitted as part of HealthPartners Total Cost of Care and Total Resource Use NQF submission.
Steve Connnelly, MD, President, Park Nicollet Health Services and Kristi Lyon, Vice President, Payer Relations

- Ms. Lori Martin on Behalf of Andrew Dorwart from HealthPartners on 3/1/17:

Stillwater Medical Group and Lakeview Hospital is an integrated, non-profit clinic and hospital system serving the eastern Twin Cities metro area and Western Wisconsin. We use HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598) measures in our system to identify opportunities to improve affordability for our patients. We support maintaining endorsement of the HealthPartners measures.
Andrew Dorwart, MD
Stillwater Medical Group President, Lakeview Hospital System CMO

- Sia Lo on Behalf of Brian Rank, MD from HealthPartners Medical Group on 3/2/17:

HealthPartners Medical Group strongly recommends for endorsement both the Total Cost of Care (#1604) and Total Resource Use (#1598) measures. For more than a decade, Total Cost of Care (TCOC) has been the top line measure of affordability for our care group. We drill down from the overall measure of TCOC to price drivers, and Total Resource Use drivers to identify opportunities for improvement. These measures have guided our improvement strategies; allowing us to focus on appropriate use of services and place of service opportunities. This has resulted in improved affordability for our patients. Our full statement of support and usability of these measures was included in the measure submission.
Nancy McClure, Chief Operating Officer and Brian Rank, MD, Executive Medical Director, and Beth Averbeck, MD, Senior Medical Director Primary Care

- Dr. Paul Kasuba from Tufts Health Plan comment on 3/3/17:

Tufts Health Plan supports endorsement of the Health Partners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. These measures have been widely adopted by many stakeholders in the health care community and have advanced the national conversation of health care affordability.
Paul Kasuba, MD SVP/CMO

- Thomas Foels from Independent Health comment on 3/5/17:

Independent Health supports endorsement of HealthPartner's Total Cost of Care (#1604 and #1598)
measures. These measures have been adopted by many stakeholders in the health care community and have advanced the national discussion on health care affordability.

- Angelo Sinopoli from Greenville Hospital System comment on 3/5/17:

Greenville Health System fully supports endorsement of the Health Partners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. These measures have been widely adopted by many stakeholders in the healthcare community and have advanced the national conversation around healthcare affordability.

Angelo Sinopoli, MD
VP, Clinical Integration, CMO

- Mr. Akinluwa Demehin, MPH from American Hospital Association comment on 3/6/17:

The American Hospital Association (AHA) recognizes the importance of total cost of care and resource use measures in helping those running health plans better understand and address opportunities to improve the value of the care provided. Therefore, we are exploring a partnership with HealthPartners to pilot use of their measures (#1604 Total Cost of Care and #1598 Total Resource Use), with the goal of using these measures with a subset of our members with health plans to help them better understand their performance. We look forward to working with HealthPartners on designing and implementing this important pilot to enhance value of care for the patients and communities our member organizations serve. Our full letter was included with the HealthPartners submission documents.

- Dr. Stephen Perkins, MD from UPMC Health Plan comment on 3/6/17:

UPMC Health Plan supports endorsement of the HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. These measures have been widely adopted by many stakeholders in the health care community and have advanced the national conversation of health care affordability.

- Koryn Y. Rubin from American Medical Association comment on 3/6/17:

Given measure 1598 and 1604 are maintenance measures, the AMA would have expected the developer, HealthPartners to have provided more information on actual performance data and how well the measures performed in the real world across different groups. The developer references all of the groups that started collecting the measure as an indicator that there is progress toward improvement, but uptake of a measure does not mean the same thing as improving performance. We, therefore, have the following concerns:

The measure submission documents state that many groups and institutions are collecting and reporting the measure under the testing and usability section, but we are only provided data from HealthPartner groups in Minnesota and Western Wisconsin. We would like for data from the first submission and anything within the last 4 years to be included and for the data to include mean, std dev, min, max, interquartile range, and scores by decile. It is also not clear to us how HealthPartners standardizes prices.

We also seek clarification on the sample size. The document states it has been tested with a minimum attributed population of 600 members, but it is not clear whether this is with each practice group or by payer or plan. The reliability testing discussion also fails to address the sample size question and the number of physicians or patient that must be attributed to a group for the measure to be considered reliable. This issue was raised as a concern when the measure underwent its last review and once, again, we request more clarity around the level of analysis and how a physician group is defined.

We also find the risk-adjustment strategy utilized for this measure insufficient. The developer utilizes the ACG system which is proprietary and groups must pay to use it. The developer states you can use others but no testing of other risk-adjustment strategies is outlined to compare the results of different tools. It would be helpful to know whether the groups that implemented the measure are all using the ACG system. If not, then it is not quite clear whether the measure produces comparable results across institutions. With the SES analysis, we do not believe the developer provided an adequate conceptual analysis or sufficient information on why they did not test one of the two factors. They first state that they looked at two factors (income and education), cite one or two articles and then they say they could only look at one- income. Therefore, we do not believe what was provided is sufficient to satisfy the SES trial requirements.

We also are concerned with the definition of primary care physician because it includes specialties such as OB/GYN that have higher intensity of services. It would also be helpful to have validity testing that includes comparisons across the different specialties that are defined as primary care physicians by the measure developer and then against all of the groups to see if it can distinguish meaningful differences and not yield inaccurate comparisons by specialty.

- Russ John Kuzel comment on 3/6/17:

SelectHealth supports endorsement of the HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598). These measures have been widely adopted by many stakeholders in the health care community and have advanced the national conversation of health care affordability.

- Sanne Jones Magnan comment on 3/6/17:

Thank you for the opportunity to share my support for the HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. With my internal medicine background and my previous leadership roles as the Minnesota Commissioner of Health and President & CEO of the Institute for Clinical Systems Improvement, I know firsthand the importance of the Triple Aim for our communities and our patients. The Total Cost of Care and Total Resource Use measures help leaders, decision-makers, and physicians identify improvement opportunities for affordability and value in our healthcare systems. The measures provide transparent information needed to drive change for better health and experience at a lower cost for our patients and communities.

## Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

**IM1. High Priority**

**IM1.1. Demonstrated High Priority Aspect of Healthcare**

Affects large numbers

High resource use

Patient/societal consequences of poor quality

Severity of illness

**IM1.2. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in IM.1.3.**In 2014, health care spending represented 17 percent of US gross domestic product (GDP); this is the largest percentage of any developed nation in the world.1 A recent survey published by the Commonwealth Fund shows that while the Affordable Care Act has expanded health care coverage, adults in the United States are much more likely to go without needed care because of cost than eleven other westernized countries.2 Consequently, affordability of care continues to be a highly discussed issue, but in spite of this, prior to 2012, there were few publically available cost or resource use measures.3,4 Aware of this issue, HealthPartners developed a measurement approach in the late nineties to increase awareness of cost of care and healthcare spending for stakeholders. Total cost reflects a mix of complicated factors including, service utilization, and negotiated prices.3 Non-condition specific cost of care and resource use measures provides valuable information on how to make health care more affordable because health plans and providers can use the data to identify areas where they can lower cost by improving resource use or shifting to less expensive resources (for example, use of a surgery center instead of a hospital where medically appropriate). Evidence supports the idea that improving use of resources and price can lead to lower costs with no loss in quality. Turbyville, et al (2011) found that medical resource use has no relationship with quality of care for diabetes. 5 Fisher, et al (2004) performed a study that showed a similar result for resource use and quality of care in Academic Medical Centers.6 The Medicare Payment Advisory Commission in a report to congress in 2006 also reported that they found no correlation between higher resource use and higher quality of care across six metropolitan statistical areas (MSAs).7

Cost of care and resource use measures can be used to support a comprehensive measurement system.8 Glass, et al call for reporting of cost and resource use in ACO models as a recommended tool to improve value, they also suggest the use of resources measurement to set targets for payment incentives, by tying payments to quality and resource use improvements.9,10 In addition, overuse of health care services has led to wide variation in health care cost and use across geographies. Studies suggest that Medicare spending would decrease by almost 30 percent if medium and high spending geographies consumed health care services comparable to that of lower spending regions.11 Experts agree that reducing overuse can make care safer and more efficient.12,13 The Resource Use Index, which controls for both cost and illness burden, can be used to identify areas of overuse in health care as well as measure targeted improvement efforts.

**IM1.3. Citations for data demonstrating high priority provided in IM.1.2**

1 The World Bank. Health expenditure, total (% of GDP). http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS?end=2014&locations=US&start=1995&view=chart

2 In a New Survey of 11 Countries, US Adults Still Struggle with Access to and Affordability of Health Care. The Commonwealth Fund. November 16, 2016. http://www.commonwealthfund.org/publications/in-the-literature/2016/nov/2016-international-health-policy-survey-of-adults

3 National Committee for Quality Assurance, Insights for Improvement - Measuring Health Care Value: Relative Resource Use, 2010, http://www.ncqa.org/portals/0/hedisqm/RRU/BI%20NCQA_RRU_Publication_FINAL.pdf

4 National Quality Forum. NQF Endorses Resource Use Measures. http://www.qualityforum.org/News_And_Resources/Press_Releases/2012/NQF_Endorses_Resource_Use_Measures.aspx

5 Turbyville, Sally E., Meredith B. Rosenthal, L. Gregory Pawlson, and Sarah Hudson Scholle, Health Plan Resource Use – Bringing Us Closer to Value-Based Decision Making, The American Journal of Managed Care, 2011. Vol. 1, no. 1, p. 68-74. Last accessed http://www.ajmc.com/journals/issue/2011/2011-1-vol17-n1/ajmc_2011jan_turbyville_68to74/P-1

6 Fisher, Elliot S., David E. Wennberg, Therese A. Stukel, and Daniel J. Gottlieb, Variations in the Longitudinal Efficiency of Academic Medical Centers, Health Affairs, 2004. doi:10.1377/hlthaff.var.19. http://content.healthaffairs.org/content/early/2004/10/07/hlthaff.var.19.short

7 Medicare Payment Advisory Committee, Report to the Congress: Increasing the Value of Medicare, 2006. http://www.medpac.gov/docs/default-source/reports/Jun06_EntireReport.pdf?sfvrsn=0

8 Fisher, Elliot S.; Shortell, Stephen M. Accountable Care Organizations: Accountable for What, to Whom and How. Journal of American Medical Association. October 20, 2010. http://jama.ama-assn.org/content/304/15/1715.full

9 Glass, David; Stensland, Jeff. Accountable Care Organizations. April 9, 2008. http://medpac.gov/docs/default-source/meeting-materials/april-2008-meeting-transcript.pdf?sfvrsn=0

10.Glass, David; Stensland, Jeff. Accountable Care Organizations. March 12, 2009. http://medpac.gov/docs/default-source/meeting-materials/march-2009-meeting-transcript.pdf?sfvrsn=0

11 Skinner, Jonathan; Fisher, Elliott. The Dartmouth Atlas. Reflections on Geographic Variation in U.S. Health Care. http://www.dartmouthatlas.org/downloads/press/Skinner_Fisher_DA_05_10.pdf

12 National Quality Forum Issue Brief. Waste Not, Want Not: The Right Care for Every Patient. June 2009. www.qualityforum.org/Publications/2009/07/Waste_Not_Want_Not_Issue_Brief.aspx

13 National Priorities and Goals. National Priorities Partnership convened by the National Quality Forum. November 2008. https://www.qualityforum.org/Setting_Priorities/NPP/National_Priorities_Partnership_Goals.aspx

IM2. **Opportunity for Improvement**

IM2.1. **Briefly explain the rationale for this measure (e.g., the benefits or improvements in performance envisioned by use of this measure)**

By measuring population based relative resource use, health plans and providers can improve the affordability of health care without sacrificing quality. HealthPartners' RUI gives provider groups valuable information on resource use and, when viewed in conjunction with quality metrics, information on the efficiency of care. The HealthPartners RUI measure is a population-based, patient-centered, total resource use measure, created with Total Care Relative Resource Values that cross all categories of health services. This is in contrast to the many, episodic based resource use measures available in the market today. Both population based and episodic based resource use measures are important and complementary but a key benefit of population based measures is helping to better understand potential overuse & underuse (e.g., although efficient at spine surgery, may be performing too many).

IM2.2. **Provide performance scores on the measure as specified** (current and over time) **at the specified level of analysis.** (This is required for endorsement maintenance. Include mean, stddev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include). **This information also will be used to address the subcriterion on improvement (U.2.1.) under Usability and Use.**

The Dartmouth Atlas has been an eye-opening look at the variation in health care spending and resource use across regions for the

Medicare population. The measurement of cost of care and resource use is as widely varied in the commercial population across geographies.1

While HealthPartners has applied the measure on the commercial population, the measure could easily be applied to other populations.

A study of the Minnesota market further highlighted the significant variation in cost and efficiency ranging from $2,400 to $4,700 PMPY. Additional findings found no relation to quality or type of practice (large, small, integrated, etc).2 These findings are

further confirmed based on HealthPartners own experience and analyses.

Existing total cost and resource use measures are largely condition or episode specific measures. Prior to 2012, there was not an existing total population cost of care measure in the market that crossed all care services.3

Based on 2015 dates of service, the multi-stakeholder community collaborative, Minnesota Community Measurement (MNCM) measured the Total Resource Use of 257 provider groups, representing 1.5 million patients receiving care. The data were sourced from the four major commercial payer in Minnesota. MNCM found that risk-adjusted medical group resource use had variation up to 55 percent, from 22% below the state average to 33% above the state average. Resource use is presented relative to the state-wide average.4

HealthPartners uses Total Care Relative Resource Values, which plots all health care services, regardless of service category on a grand linear scale. Therefore, resource use can be compared across service categories where services are relative to each other. Resource use indices can be drilled down to the service category or condition to help identify areas of opportunity, especially when paired with utilization data.

IM2.3. **If no or limited performance data on the measure as specified is reported in IM.2.2., then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

1.Dartmouth Atlas. http://www.dartmouthatlas.org/

2.Kralewski, John E, Dowd, Bryan E, Xu, Yi (Wendy). Differences in the Cost of Health Care Provided by Group Practices in Minnesota. February 2011. Minnesota Medicine. http://www.minnesotamedicine.com/tabid/3678/Default.aspx

3.Berwick, Donald M., Nolan, Thomas W., Whittington, John, The Triple Aim: Care, Health and Cost. Health Affairs, May/June 2008.

doi: 10.1377/hlthaff.27.3.759. http://content.healthaffairs.org/content/27/3/759.full?sid=f3d381e8-76ef-415f-9080-de97c1273fa6

4. Minnesota Community Measurement. 2016 Cost and Utilization Report: Average Cost per procedure, Total Cost of Care Relative Resource Use, Utilization. http://mncm.org/wp-content/uploads/2016/12/16CostUtilityReport.pdf

IM2.4. **Provide disparities data from the measure as specified** (current and over time) **by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** (This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) **This information also will be used to address the subcriterion on improvement (U.2.1.) under Usability and Use.**

As previously described in this application, the measure is being submitted for a commercially insured population. Therefore performance by insurance status is not applicable because the population is all commercially insured. The clinical risk adjustment process described in 2b4.3 describes how age and gender are accounted for in the methodology and no additional measure performance was tested because this is not how they are being used. That said, in looking at single specialty obstetric and pediatric groups, we see a uniformly distributed result across our network performance and these groups are not clustered, which demonstrates results are not biased against age or gender. Additionally, this demonstrates the clinical risk adjustment is working effectively. The measure is used as a population-based method primarily for payment, benefit design, transparency and improvement.

After applying clinical risk adjustment, socioeconomic testing was conducted that considered income and education status as potential factors beyond those already adjusted for.

Model Results
1% Income Increase:
Total Reimbursement $(0.13)
Resource Use $0.16
Price $(0.28)

1% ACG Increase:
Total Reimbursement $4.22
Resource Use $4.34
Price $(0.07)

Commercial vs. Medicaid Membership:
Total Reimbursement $133.28
Resource Use $(75.24)
Price $205.36

Resource Use Endorsed Measure R2 = 0.5788
Resource Use Endorsed Measure + Income R2 = 0.5792

Using Census tract data, a 1% increase in income resulted in a $0.13 decrease in total reimbursement, a $0.16 increase in resource use, and $0.28 decrease in price. The results highlight how significantly more the ACG score (clinical risk adjustment) and insurance product impact both the cost and resource use measures. For frame of reference, on average for the Midwest market, the total spend for a member per month (PMPM) is $400. The results of the evaluation show that a 1% increase in risk score accounts for a $4.22 or roughly 1% increase in PMPM.

Product also contributed significantly with there being a $133 dollar difference in cost between commercial and Medicaid. The variation in resource use was much less, however, still significant with Medicaid covered members utilizing $75 more dollars of resources. The fact that Medicaid's cost per service is approximately half that of commercial rates drives the differences between the TCOC and Resource Use results. The R2 results further emphasize that ACG score and insurance type are the main drivers of cost and resource use variation and income does not provide any additional predictive power.

Methodology and testing results can be found here:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

IM2.5. **If no or limited data on disparities from the measure as specified is reported in IM.2.4., then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Not applicable

IM3. **Measure Intent**

IM3.1. **Describe intent of the measure and its components/ Rationale (including any citations) for analyzing variation in resource use in this way.**

Key considerations when constructing the measure:

- The purpose of population-based measurement is to better understand overuse, underuse, and person-centered management and accountability
- Population based-measurement nicely complements condition and episode-base measures; combined they depict a complete picture of a provider's resource use
- Risk adjustment is a critical component to the measure to allow for fair comparisons
- Use this measure as part of a Triple-aim approach where Total Resource Use is a complement to total cost of care, quality and patient experience.
- Removing price via Total Care Relative Resource Values (TCRRVs) allows for a clear picture of resource use opportunities.
- The Resource Use Index measure, when used with a Total Cost of Care measure, will help to better understand cost and resource use opportunities.

## Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

**Specifications** The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Non-Condition Specific** *(check all the areas that apply):*
Care Coordination
Safety : Overuse

De.7. **Care Setting** *(Select all the settings for which the measure is specified and tested):*
Ambulatory Surgery Center
Behavioral Health : Inpatient
Behavioral Health : Outpatient
Birthing Center
Clinician Office/Clinic
Dialysis Facility
Emergency Department
Emergency Medical Services/Ambulance
Home Health
Hospice
Hospital
Hospital : Acute Care Facility
Hospital : Critical Care
Imaging Facility
Inpatient Rehabilitation Facility
Laboratory
Long Term Acute Care
Nursing Home / SNF
Other:All care settings included
Outpatient Rehabilitation
Pharmacy
Urgent Care - Ambulatory

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
For purposes of resubmission please use the following link to view materials including updated measure specifications: www.healthpartners.com/tcoc-documents For reference, currently endorsed measure materials reside here: www.healthpartners.com/tcoc

S.2. **Type of resource use measure** *(Select the most relevant)*
Per capita (population- or patient-based)

S.3. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED):*
Clinician : Group/Practice, Population : Community, County or City

S.4. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*

S.5. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
*If other, please describe in S.5.1.*
Claims (Only)

S.5.1. **Data Source or Collection Instrument** *(Identify the specific data source or data collection instrument, e.g. name of database, clinical registry, collection instrument, etc.)*

- Users administrative claims data base
- Risk Adjustment Tool, Johns Hopkins ACG System
- Standardized costing code table, Total Care Relative Resource Values (TCRRV) specification provided

S.5.2. **Data Source or Collection Instrument Reference** *(available at measure-specific Web page URL identified in S.1 OR in the file attached here) (Save file as: S_5_2_DataSourceReference)*

S.6. **Data Dictionary or Code Table** *(Please provide a web page URL or attachment if exceeds 2 pages. NQF strongly prefers URLs. Attach documents only if they are not available on a web page.)*

***Data Dictionary:***

URL:

Please supply the username and password:

Attachment:

***Code Table:***

URL:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188112.pdf

Please supply the username and password:

Attachment:

**Construction Logic**

S.7.1. **Brief Description of Construction Logic**

If applicable, summarize the general approach or methodology to the measure construction. This is most relevant to measures that are part of or rely on the execution of a measure system or applies to multiple measures.

The measure examines total resource use of a commercial population for a given measurement year (e.g. January 1 and December 31), for all members eligible for the measure.

S.7.2. **Construction Logic** *(Detail logic steps used to cluster, group or assign claims beyond those associated with the measure's clinical logic.)*

- All claims included in the measure have a date of service in the measurement year (e.g. between January 1 and December 31)
- Members have a minimum 9 months enrollment in the measurement year
- Commercial population only
- Attribution
- Costing Method – Total Care Relative Resource Values TCCRVs
- Risk Adjustment

S.7.2a. **CONSTRUCTION LOGIC ATTACHMENT or URL:** If needed, attach supplemental documentation (Save file as: S_7_2_Construction_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.

URL: https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_187908.pdf

Please supply the username and password:

Attachment:

S.7.3. **Concurrency of clinical events, measure redundancy or overlap, disease interactions** *(Detail the method used for identifying concurrent clinical events, how to manage them, and provide the rationale for this methodology.)*
We do not provide specifications for concurrency of clinical events.

Not applicable. This is a population-based measure that applies to all care settings and conditions.

S.7.4. **Complementary services** *(Detail how complementary services have been linked to the measure and provide rationale for this methodology.)*
We do not provide specifications for linking complementary services.

Not applicable. This is a population-based measure that applies to all care settings and conditions.

S.7.5. **Clinical hierarchies** *(Detail the hierarchy of codes or condition groups used and provide rationale for this methodology.)*
We do not provide specifications for clinical hierarchies.

S.7.6. **Missing Data** *(Detail steps associated with missing data and provide rationale for this methodology (e.g., any statistical techniques to impute missing data)*
We do not provide measure specifications or guidelines for missing data :

In the instances where members have pharmacy benefit carve-outs the following methodology is applied.

The Resource Use measure accounts for members that have their pharmacy benefit carved out by using the members that have pharmacy coverage as a proxy. This technique allows for members without pharmacy coverage to be included in the medical portion of the total resource use with their pharmacy TCRRVs reflecting the provider's risk adjusted pharmacy TCRRVs from those covered. The measures separate the total resource use into medical and pharmacy and only includes the members with pharmacy coverage into the TCRRV PMPM calculation for pharmacy. The total TCRRV PMPM for a provider group is then calculated by adding the medical TCRRV PMPM to the pharmacy TCRRV PMPM: Total TCRRV PMPM = (Medical TCRRVs / Medical MMs) + (Pharmacy TCRRVs / Pharmacy MMs). MM = member months.

HealthPartners' data includes all medical and mental health care. It also includes the majority of pharmacy claims with the exception of some carveouts. The methodology described above was used for testing. If users have additional carve-outs (e.g., mental health) the lowest common denominator principle (i.e. for any given user if their data includes a carve-out for one their method must apply a carve-out for all) needs to be applied to ensure providers are evaluated fairly. This would require all services that are carved out of one segment of input data to be carved out of the measure for all segments of input data and all input components of the measure (e.g. TCRRV PMPMs, attribution, and risk adjustment).

S.7.7. **Resource Use Service Categories (Units) (Select all categories that apply)**

Inpatient services: Inpatient facility services

Inpatient services: Evaluation and management

Inpatient services: Procedures and surgeries

Inpatient services: Imaging and diagnostic

Inpatient services: Lab services

Inpatient services: Admissions/discharges

Inpatient services: Labor (hours, FTE, etc.)

Other inpatient services

Ambulatory services: Outpatient facility services

Ambulatory services: Emergency Department

Ambulatory services: Pharmacy

Ambulatory services: Evaluation and management

Ambulatory services: Procedures and surgeries

Ambulatory services: Imaging and diagnostic

Ambulatory services: Lab services

Ambulatory services: Labor (hours, FTE, etc.)

Other ambulatory services

Durable Medical Equipment (DME)

Other services not listed

All care is included

All care is included

All care is included

S.7.8. **Identification of Resource Use Service Categories (Units)**
*(For each of the resource use service categories selected above, provide the rationale for their selection and detail the method or algorithms to identify resource units, including codes, logic and definitions.)*
The Total Resource Use considers 100% of health care services in the Resource Use Index and is calculated on a risk-adjusted paid per member per month basis as well as benchmarked to a peer group. The Total Care Relative Resource Values (TCRRVTM) is inclusive of both plan and member liability. Detailed identification of units is available in the Total Care Relative Resource Value White Paper.
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_039627.pdf

S.7.8a. **If needed, provide supplemental resource use service category specifications in either URL (preferred) or as an attachment (Save file as S.7.8a_RU_Service_Categories):**
URL:

Please supply the username and password:

Attachment:

**Clinical Logic**

S.8.1. **Brief Description of Clinical Logic** (Briefly describe your clinical logic approach including clinical topic area, whether or not your account for comorbid and interactions, clinical hierarchies, clinical severity levels and concurrency of clinical events.)
Not applicable. This is a population-based measure that applies to all care settings and conditions.

S.8.2. **Clinical Logic** *(Detail any clustering and the assignment of codes, including the grouping methodology, the assignment algorithm, and relevant codes for these methodologies.)*
Not applicable. This is a population-based measure that applies to all care settings and conditions.

S.8.3. **Evidence to Support Clinical Logic Described in S.8.2** *Describe the rationale, citing evidence to support the grouping of clinical conditions in the measurement population(s) and the intent of the measure (as described in IM3)*
Not applicable. This is a population-based measure that applies to all care settings and conditions.

S.8.3a. **CLINICAL LOGIC ATTACHMENT or URL: If needed, attach supplemental documentation (Save file as: S_8_3a_Clinical_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.**
URL:

Please supply the username and password:

Attachment:

S.8.4. **Measure Trigger and End mechanisms** *(Detail the measure's trigger and end mechanisms and provide rationale for this methodology)*

All claims dates of service in the measurement year (e.g. January 1 – December 31)

S.8.5. **Clinical severity levels** *(Detail the method used for assigning severity level and provide rationale for this methodology)*

We do not provide specifications for clinical severity levels.

This is accounted for in application of risk adjustment, Johns Hopkins, ACG System

S.8.6. **Comorbid and interactions** *(Detail the treatment of co-morbidities and disease interactions and provide rationale for this methodology.)*

We do not provide specifications for co-morbidies and disease interactions.

This is accounted for in application of risk adjustment, Johns Hopkins, ACG System

**Adjustments for Comparability**

S.9.1. **Inclusion and Exclusion Criteria** *Detail initial inclusion/exclusion criteria and data preparation steps (related to clinical exclusions, claim-line or other data quality, data validation, e.g. truncation or removal of low or high dollar claim, exclusion of ESRD patients)*

We do not provide measure specifications or guidelines for data inclusion criteria :
The HealthPartners' Total Resource Use measure is a full population-based measure, with members under age 1, members 65+ and members with less than 9 months of enrollment excluded to ensure an accurate risk assessment is made on the population.

- Members over age 64
- Members under age 1
- Member enrollment less than nine months during the one year measurement time window
- TCRRVs per member up to 125,000 are included; TCRRVs per member above 125,000 are excluded (truncated)

• Administrative claims covering all categories of health care services: professional, facility inpatient and outpatient, pharmacy, lab, radiology and any other ancillary healthcare services are included.
• Johns Hopkins ACG System for risk adjustment
• Membership eligibility, identifier and number of months during the measurement period the member was eligible (member months)
The following should be reviewed prior to beginning implementation of the Total Resource Use measure to ensure data comparability:
• Consistent population of primary and secondary claims diagnosis. Population prevalence to ensure reasonable/completeness of
disease; primary and secondary diagnosis are consistently populated (e.g., diagnosis 1 - 4)
• Data elements are populated within reasonable tolerances and thresholds (e.g., expected CPT ranges, expected allowed amount
ranges, expected units ranges)
• All service categories are available and appropriately represented (e.g., inpatient, pharmacy, outpatient and professional)
• Peer group/case-mix need to be comparable
• Risk adjustment weight and application must be in sync (e.g. truncation threshold values)
It is recommended that further reliability and validity testing be conducted if the user varies from the "Technical Guidelines" provided. Examples include:
• The user implements the measure with less than 600 members attributed to a provider
• The user applies a different unit of evaluation, such as an employer group, condition or community rather than a provider
• The user employs an alternative attribution algorithm or risk adjustment tool

Paid medical and pharmacy administrative claims for the measurement year (e.g. between January 1 and December 31), allowing
for three months of run out for claims lag.

In the instances where members have pharmacy benefit carve-outs the following methodology is applied.
The Resource Use measure accounts for members that have their pharmacy benefit carved out by using the members that have pharmacy coverage as a proxy. This technique allows for members without pharmacy coverage to be included in the medical portion of the total resource use with their pharmacy TCRRVs reflecting the provider's risk adjusted pharmacy TCRRVs from those covered. The measures separate the total resource use into medical and pharmacy and only includes the members with

pharmacy coverage into the TCRRV PMPM calculation for pharmacy. The total TCRRV PMPM for a provider group is then calculated by adding the medical TCRRV PMPM to the pharmacy TCRRV PMPM: Total TCRRV PMPM = (Medical TCRRVs / Medical MMs) + (Pharmacy TCRRVs / Pharmacy MMs). MM = member months.

HealthPartners' data includes all medical and mental health care. It also includes the majority of pharmacy claims with the exception of some carveouts. The methodology described above was used for testing. If users have additional carve-outs (e.g., mental health) the lowest common denominator principle (i.e. for any given user if their data includes a carve-out for one their method must apply a carve-out for all) needs to be applied to ensure providers are evaluated fairly. This would require all services that are carved out of one segment of input data to be carved out of the measure for all segments of input data and all input components of the measure (e.g. TCRRV PMPMs, attribution, and risk adjustment).

S.9.2. **Risk Adjustment Type** (Select type)
Statistical risk model
If other:

S.9.3. **Statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables.)*
For Total Resource Use measurement, risk adjustment is performed using Adjusted Clinical Groups (ACG) developed by Johns Hopkins University. The Johns Hopkins ACG® System has the distinction of being developed, tested and supported by a world renowned academic and medical research institution, The Johns Hopkins University. The academic home of the ACG System allows for an unparalleled openness to the method. Each component of the system is exposed to the user which allows the system to be easily adapted to unique local circumstances and applications. The ACG methodology is subject to continuous critical review and testing by a team of distinguished health services researchers led by Dr. Jonathan Weiner. This transparency and academic credibility is critical when trying to disseminate risk information to providers and purchasers of healthcare. Attributed members are assigned a risk score based on diagnoses on claims from the performance measurement period, as well as member age and gender. The Society of Actuaries Accuracy of Claims-Based Risk Scoring Models (2016) findings suggest other comparable risk groupers are available and would need to be tested for reliability and validity of that risk grouper.
https://www.soa.org/Files/Research/research-2016-accuracy-claims-based-risk-scoring-models.pdf

For the purpose of this application, this measure has been tested using the Johns Hopkins University developed Adjusted Clinical Groups (ACG System).

http://acg.jhsph.org/

Technical Paper: https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/dev_057425.pdf
Risk Adjustment Specifications
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/dev_057913.pdf
ACG Technical Guide
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf

See measure testing attachment for more information on the statistical risk model method and variables.

S.9.4. **Detailed Risk Model Specifications** *available at measure-specific Web page URL identified in S.1 OR in attached data dictionary/code list Excel or csv file.*
Available at measure-specific web page URL identified in S.1

S.9.5. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets)*
Measures are adjusted for clinical risk and limited to the commercial population.

S.9.6. **Costing method**
Detail the costing method including the source of cost information, steps to capture, apply or estimate cost information, and provide rationale for this methodology.
Standardized pricing
Total Resource Use measure uses the Total Care Relative Resource Values (TCRRVs). TCRRVs are a grand linear scale of relative values designed to evaluate resource use across all types of medical services, procedures and places of service. The values are

independent of price and can be used to evaluate providers, hospitals, physicians and health plans against their peers on their efficiency of resource use in treating like conditions.

General Overview of Application:

The TCRRVs are applied at the procedure level for each component of care with the exception of inpatient, which is applied at the full admission level. There is a TCRRV lookup table for each component of care where each claim's procedure is matched with the corresponding value. The TCRRV weights that are applied to the claim is tested for accuracy and a total TCRRV is calculated.

Detail development:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_039627.pdf

Sample TCRRV table:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188112.pdf

S.10. **Type of score***(Select the most relevant):*
Ratio
Other (specify):
If other: https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/dev_057910.pdf see page 9
Attachment:

S.11. **Interpretation of Score** *(Classifies interpretation of a ratio score(s) according to whether higher or lower resource use amounts is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score, etc.)*
A provider Total Resource Use Index (RUI) of 1.10 equates to 10% higher risk adjusted resource use. Similarly, a provider RUI score of 0.90 equates to 10% less paid risk adjusted resource use.
A score of 1.0 is equivalent to the peer group average.

S.12. **Detail Score Estimation** *(Detail steps to estimate measure score.)*
There is no estimation in the Total Resource Use Measure. The actual result is calculated as follows:
Resource Use Index (RUI):
Numerator: Total Resource PMPM = (Total Medical TCRRV / Medical Member Months) + (Total Pharmacy TCRRV / Pharmacy Member Months)

Denominator: Average Risk Score - the medical claims data is submitted through the Johns Hopkins ACG Risk Grouper which generates a relative risk score for each member. That risk score is then multiplied by the number of months a member has been enrolled creating a risk weight. The risk weights are then summed to the desired level of measurement (e.g., provider group) and divided by the total sum of the desired level's member months creating a member month weighted Average Risk Score.

ACG Adjusted Total Resource Use PMPM = Total Resource Use PMPM / ACG Risk Score
Resource Use Index = Provider ACG Adjusted Total Resource Use PMPM / Peer Group ACG Adjusted Total Resource Use PMPM

**Reporting Guidelines**
This section is optional and will be available for users of the measure as guidance for implementation and reporting.

S.13.1. **Describe discriminating results approach**

Detail methods for discriminating differences (reporting with descriptive statistics--e.g., distribution, confidence intervals).
This is a full population-based measure, therefore, confidence intervals are not applicable. The results can be analyzed by percentile, percent from mean, standard deviation and clustering methods, this is dependent upon the business application of the
measure.

A provider Total Resource Use Index (RUI) of 1.10 equates to 10% higher risk adjusted resource use. Similarly, a provider RUI score of 0.90 equates to 10% less paid risk adjusted resource use.
A score of 1.00 is equivalent to the peer group average.

S.13.2. **Detail attribution approach**

Detail the attribution rules used for attributing resources/costs to providers (e.g., a proportion of total measure cost or frequency of visits during the measure's measurement period) and provide rationale for this methodology.

There are three main options to include members in the Total Resource Use measure, by geographic region or defined population,  assignment of members to a responsible party or attribution of members to a responsible party. Each option will require a different approach for assigning members to the responsible party or unit of measurement. In all cases the measure exclusions will still need to be applied (i.e., age 1-64, commercial members, enrolled a minimum of 9 months).

- Population – categorize members by geographic areas such as state, county or zip code or in a defined population, for example an employer group.
- Assignment – each member is assigned to a responsible provider group regardless of where he or she received care
- Attribution – in instances where members are not assigned a responsible provider and are able to receive care at any provider (an open access market), applying an Attribution Algorithm will be necessary to identify the provider that is responsible for managing the patient's care. There are several options for defining the responsible provider, but in general the provider that sees the patient the most often is attributed the member

For purposes of testing, a commonly used Attribution Algorithm for an open access market was applied - plurality model, using most recent visit as a tie breaker:

- Include all professional claims experience claims data in a twelve month measurement period (e.g. January 1 – December 31)
- Exclude all services that are not office based
- Exclude convenience care clinic visits
- Exclude all providers that are not a physician, physician assistant or nurse practitioner
- Assign each service line a specialty based on the servicing physician's practicing specialty or credential specialty if practicing specialty is not available
- Include only the following specialties: Family Medicine, Internal Medicine, Pediatrics, Geriatrics, OB/GYN

A real-case example of using our Total Resource Use measure at a population level:

- The Network for Regional Healthcare Improvement (NRHI), a licensee of HealthPartners' measures, released their first report in their Getting to Affordability initiative in November 2016 called "From Claims to Clarity:  Deriving Actionable Healthcare Cost Benchmarks from Aggregated Commercial Claims Data".  This report compared commercially insured health plan members across five geographic regions.  When looking at the population level (regional geographic view), all members with commercial insurance were included, yet when the individual regions were evaluated at the provider group practice level, they used attribution to only include patients who had a visit with a primary care provider.

Technical specifications:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_187908.pdf

HealthPartners has studied various attribution methods, our findings are located here: HealthPartners Attribution Technical Paper
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_031064.pdf

S.13.3. **Identify and define peer group**

Identify the peer group and detail how peer group is identified and provide rationale for this methodology.

The peer group can be applied by market, region or national with the following criteria:

- Provider Specialties include: Internal Medicine, Family Medicine, Pediatrics, Geriatrics and OB/GYN
- Provider Types include: Physician, Physician Assistant, Nurse Practitioner

S.13.4. **Sample size**

Detail the sample size requirements for reporting measure results.

This measure has been tested for a minimum attributed member population of 600 members, this number is aligned with over 80+ community-based quality and patient experience measures in the market tested. We recommend further reliability and validity testing if a threshold less than 600 attributed members is used.

S.13.5. **Define benchmarking and comparative estimates**

Detail steps to produce benchmarking and comparative estimates and provide rationale for this methodology.

The Resource use measure is relative to a benchmark or peer group of the user's choice. This can be a group of members, providers, geographic regions or any grouping of member data. The idea is that the Resource Use measure will return a value that will be relative to the peer group average (e.g., 1.10 = 10% higher than the peer group average).

The peer group average is set as the benchmark and a provider's Total Resource Use ACG Adjusted PMPM is indexed against the peer group average. The Peer Group average is calculated in the same manner as an individual provider:

Resource Use (RUI):
Numerator: Peer Group Total Resource Use PMPM = (Peer Group Total Medical TCRRV/ Peer Group Medical Member Months) + (Peer Group Total Pharmacy TCRRV / Peer Group Pharmacy Member Months)

Denominator: Peer Group ACG Risk Score

Peer Group ACG Adjusted Total Resource Use PMPM = Peer Group Total Resource Use PMPM / Peer Group ACG Risk Score

Resource Use Index: RUI = Provider ACG Adjusted Total Resource Use PMPM / Peer Group ACG Adjusted Total Resource Use PMPM

**Validity – See attached Measure Testing Submission Form**

**SA.1. Attach measure testing form**
NQF_testing_attachment_Total_Resource_Use_1598_021517-636227633638474421.docx

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

**Measure Number** (*if previously endorsed*)**:** 1598
**Measure Title**: Total Resource Use Population-based PMPM Index
**Date of Submission**: 12/1/2016
**Type of Measure:**

| | |
|---|---|
| ☐ Outcome (*including PRO-PM*) | ☐ Composite – ***STOP – use composite testing form*** |
| ☐ Intermediate Clinical Outcome | X Cost/resource |
| ☐ Process | ☐ Efficiency |
| ☐ Structure | |

**Instructions**

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***

- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

**Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.**

**2a2. Reliability testing** [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

**2b2. Validity testing** [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

**2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

**AND**

If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the

exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

**2b4. For outcome measures and other measures when indicated** (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14],[15] and has demonstrated adequate discrimination and calibration

**OR**

- rationale/data support no risk adjustment/ stratification.

**2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

**OR**

there is evidence of overall less-than-optimal performance.

**2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results**.

**2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

**Notes**

**10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

**11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

**12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

**13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

**14.** Risk factors that influence outcomes should not be specified as exclusions

**15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

HealthPartners has developed a Total Cost of Care (TCOC) measure and a Total Resource Use measure. The two measures use the same measurement criteria except for the costing method. While the measures can be used independently, when used together they provide a comprehensive evaluation of cost and further identify opportunities to improve affordability. TCOC measure is a combination of resource use and price and measures the cost effectiveness of managing a population. Total Resource Use measure removes price and measures the frequency and intensity of services.

Because Resource Use is a component of Total Cost of Care, the two measures are complementary to each other, therefore the two measures are tested and evaluated together for reliability and validity, also increasing efficiency of testing by the measure developer. References to both measures are included in the links to technical papers and table of results found throughout the attachment.

Note: Information from prior submission in 2012 is included in *gray italic font* within the body of the form. Methodology used for testing remains the same as prior submission. Results from prior testing are included as a packaged PDF of technical papers within Appendix A. The packaged reports provide a complete analytical pathway with context and reasoning to conclude the measure is reliable and valid.

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for <u>all</u> the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

| **Measure Specified to Use Data From:**<br>**(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
|---|---|
| ☐ abstracted from paper record | ☐ abstracted from paper record |
| x☐ administrative claims | x☐ administrative claims |
| ☐ clinical database/registry | ☐ clinical database/registry |
| ☐ abstracted from electronic health record | ☐ abstracted from electronic health record |
| ☐ eMeasure (HQMF) implemented in EHRs | ☐ eMeasure (HQMF) implemented in EHRs |
| ☐ other: | ☐ other: |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Commercial administrative claims
Medicaid administrative claims were used in addition to commercial claims for purposes of socio-economic status (SES) testing.

**1.3. What are the dates of the data used in testing**?

2014, 2015 dates of service for validity testing
2015 dates of service for reliability testing
2015 dates of service for SES testing

*Prior submission: 2007, 2008, 2009 dates of service*

**1.4. What levels of analysis were tested**? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

| **Measure Specified to Measure Performance of:**<br>**(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
|---|---|
| ☐ individual clinician | ☐ individual clinician |
| x☐ group/practice | x☐ group/practice |
| ☐ hospital/facility/agency | ☐ hospital/facility/agency |
| x☐ health plan | ☐ health plan |
| ☐ other: | ☐ other: |

**1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities* *included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

HealthPartners primary care network (Minnesota and western Wisconsin) consists of 66 individual provider groups that have 850 clinic sites. Provider group size vary from 600 to a few large systems with 40,000+ members.

*Prior submission: HealthPartners' primary care Twin Cities metro area providers as per the specifications of the measure for the calendar years of 2007, 2008 and 2009. HealthPartners primary care metro network consists of 19 individual providers that have 223 (2007) 232 (2008) and 229 (2009) clinic sites.*

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

This is a population-based measure that applies to all care settings and conditions using HealthPartners health plan's full book of business. The total membership of the primary care attributed network is over 530,000 members in 2015.

*Prior submission: The total membership of the primary care attributed metro network membership grew slightly over the three year period: 268,912 (2007), 272,491 (2008) 303,638 (2009).*

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.
Reliability and Validity testing use the same population and underlying data. The SES testing also includes the Medicaid population.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

The Total Resource Use measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). The measure is also limited to commercial only. Socioeconomic testing was conducted that considered income and education status as potential factors beyond those already adjusted for.

## 2a2. RELIABILITY TESTING

***Note****: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)

□ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

x□ **Performance measure score** (e.g., *signal-to-noise analysis*)

*Prior submission: Please see Appendix A (page 8) for reliability testing results from prior submission. The method of testing (bootstrapping and 90% random sample) used for current resubmission is the same methodology used in prior submission.*

**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Overview of Analysis

Resource Use Index (RUI) is a measure of a provider's effectiveness of managing their primary care attributed population across the care continuum. The RUI measure was applied to HealthPartners primary care providers as per the measure specifications and results were calculated for 2015.

The reliability testing demonstrates the repeatability of producing the same results a high proportion of the time. To measure the reliability of the RUI measure the actual results were compared to the results calculated by two sampling methods, bootstrapping and a 90% random sample.

These methods were chosen as they represent the measure intent, which is that the RUI measure represents providers' average resource use across their population. Since the measure is aggregated to the provider group level, evaluation of member level variability is not necessary.

In the bootstrapping method members that were attributed to a provider group were randomly selected with replacement. This method artificially creates variation around a provider group's total resource use as each randomly selected iteration (sample populations) does not truly represent the provider's case mix of patients. What this method does however is give an indication as to the repeatability of the measure by comparing how closely the actual resource use measure is to the bootstrapped averages.

In the 90% random sample method, the members that were attributed to a provider group were randomly sampled at the 90% membership level without replacement. This technique was employed to create variation within a provider group by leveraging their own population and controlling for the patient case mix variation that is introduced when random sampling is employed.

Methodology

To perform the bootstrap, the SAS PROC SURVEYSELECT procedure with the Unrestricted Random Sample option for full replacement was utilized to create a series of random samples for each provider group being measured. Full replacement means that one observation is drawn at random, recorded, and then placed back into the data pool so that it can be drawn again if randomly selected. The numbers of records sampled are drawn such that the samples created are the same size as the original number of attributed members for the provider group. In this way, it is theoretically possible (although virtually improbable) to produce a sample of size n that could consist of the same record drawn n times in a row. This sample process was performed 500 times for each provider group being analyzed, to produce 500 sets of risk-adjusted Total Resource Use results for each provider in the analysis.

Once the 500 samples were created for each provider group, the total resource use of each sample for each provider group was compared to the network average to produce a risk adjusted index. The mean Total Resource Use (RUI) from these 500 iterations was computed and compared to the Actual Resource Use Index (RUI) index for each provider group.

In the second method, 90% of attributed provider group members were randomly selected, without replacement. A 90% sample was used despite having the full health plan provider population, as a concession to provider claims that errors in administrative data may not allow for a perfect 100% representation of their population. The sampling process was performed using the SAS PROC SURVEYSELECT procedure with the Simple Random Sample (SRS) option. This method allows for each attributed member to be selected only one time until 90% of the total provider population has been reached.

The 90% sampling process was repeated 500 times for each provider group analyzed. Attributed members' total resource use was aggregated in each sample to produce 500 Total Resource Use Index results for each provider group. The mean of the sampled Total Resource Use index was calculated for each provider group and compared to the Actual RUI index for each provider group.

The bootstrap results should indicate that the within provider RUI variation is significantly less than the between provider variation.

Reliability Paper includes the same method of testing described above:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188105.pdf

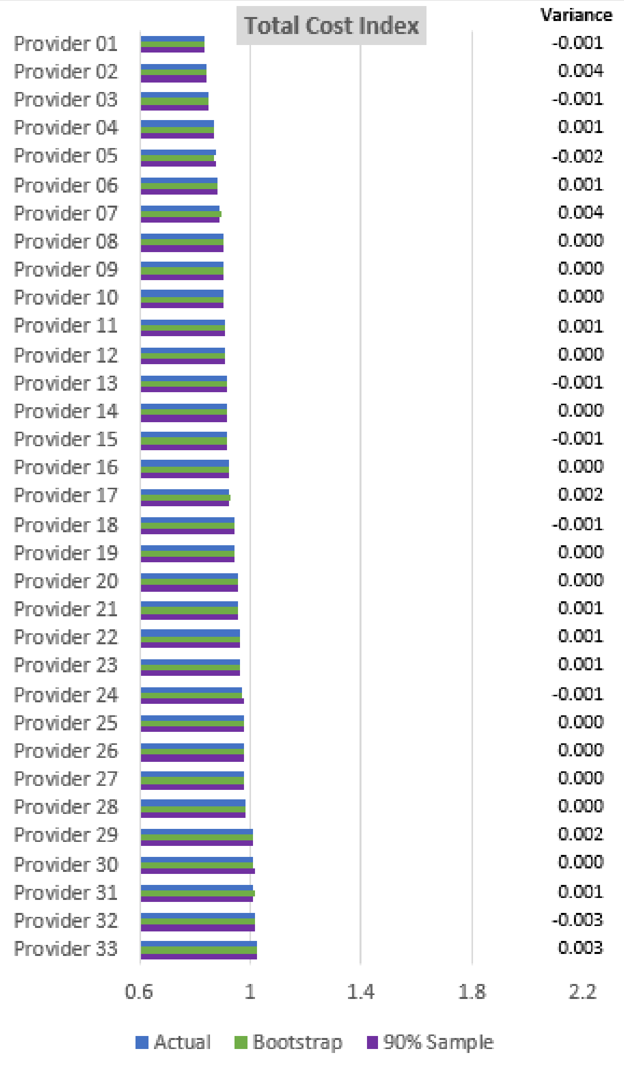
**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

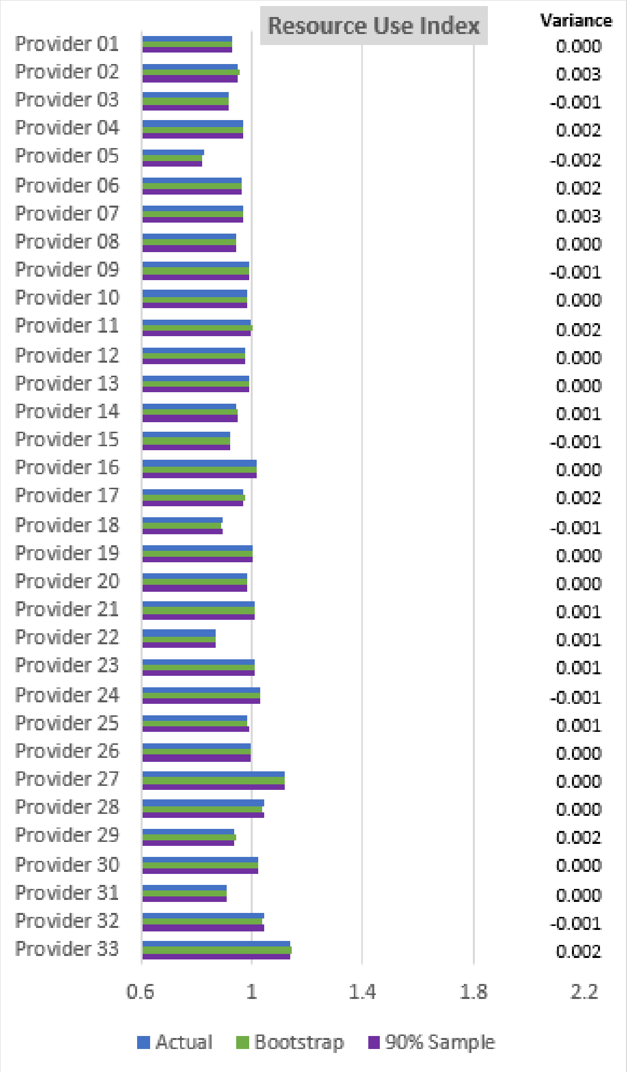
- The variances from Actual RUI ranged from -0.0036 to 0.0065 in the bootstrap to -0.0020 to 0.0015 in the 90% sample.

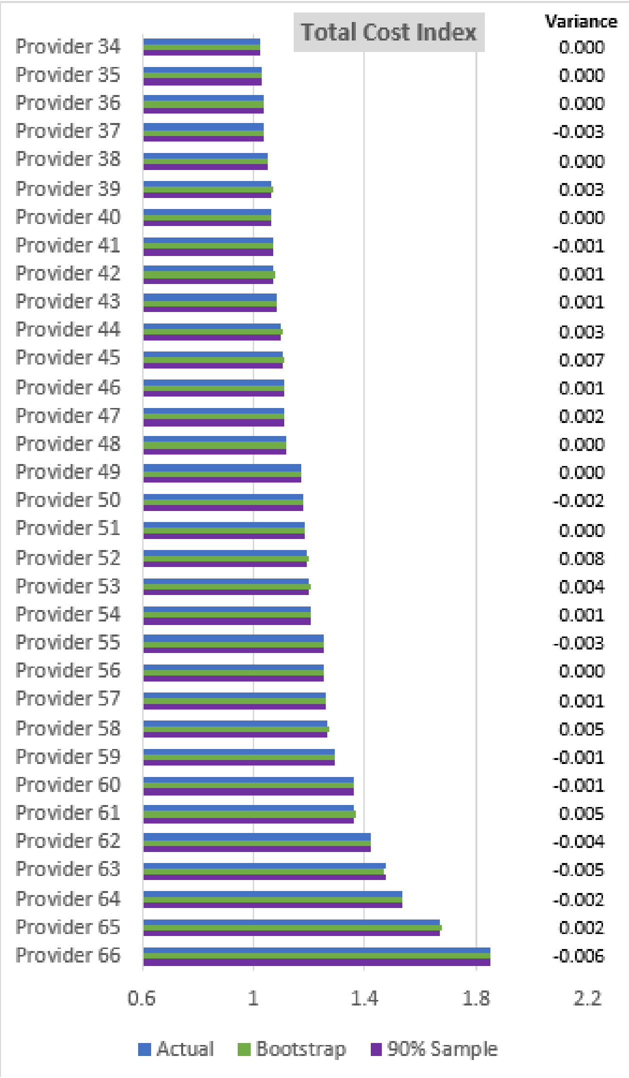
The mean Total Resource Use results from the bootstrap and 90% samples compared to the actual RUI results for each provider group are displayed on the charts on the following pages. The variance between the actual RUI to the bootstrap results is shown on the far right of each chart. The RUI charts are sorted in ascending order by Total Cost Index. See Reliability Paper for detailed results.

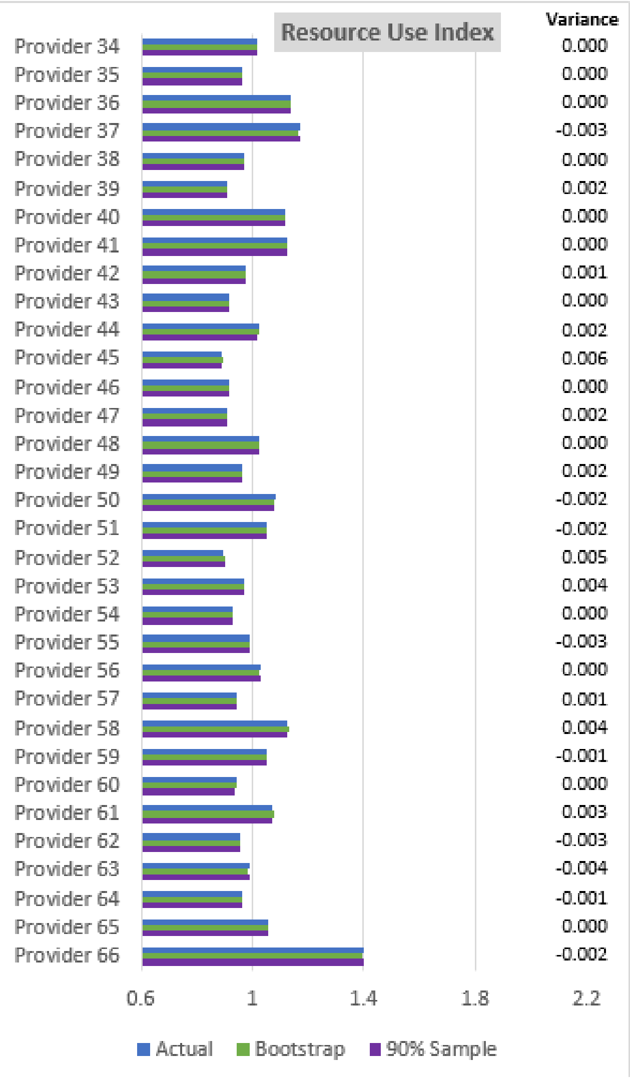
Reliability Paper describes the results of testing in detail:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188105.pdf

*Prior submission: Please see Appendix A (page 8) for prior submission results.*


Total Cost Index
Variance
Provider 01 -0.001
Provider 02 0.004
Provider 03 -0.001
Provider 04 0.001
Provider 05 -0.002
Provider 06 0.001
Provider 07 0.004
Provider 08 0.000
Provider 09 0.000
Provider 10 0.000
Provider 11 0.001
Provider 12 0.000
Provider 13 -0.001
Provider 14 0.000
Provider 15 -0.001
Provider 16 0.000
Provider 17 0.002
Provider 18 -0.001
Provider 19 0.000
Provider 20 0.000
Provider 21 0.001
Provider 22 0.001
Provider 23 0.001
Provider 24 -0.001
Provider 25 0.000
Provider 26 0.000
Provider 27 0.000
Provider 28 0.000
Provider 29 0.002
Provider 30 0.000
Provider 31 0.001
Provider 32 -0.003
Provider 33 0.003
0.6
1
1.4
1.8
2.2
Actual
Bootstrap
90% Sample


Resource Use Index
Variance
Provider 01 0.000
Provider 02 0.003
Provider 03 -0.001
Provider 04 0.002
Provider 05 -0.002
Provider 06 0.002
Provider 07 0.003
Provider 08 0.000
Provider 09 -0.001
Provider 10 0.000
Provider 11 0.002
Provider 12 0.000
Provider 13 0.000
Provider 14 0.001
Provider 15 -0.001
Provider 16 0.000
Provider 17 0.002
Provider 18 -0.001
Provider 19 0.000
Provider 20 0.000
Provider 21 0.001
Provider 22 0.001
Provider 23 0.001
Provider 24 -0.001
Provider 25 0.001
Provider 26 0.000
Provider 27 0.000
Provider 28 0.000
Provider 29 0.002
Provider 30 0.000
Provider 31 0.000
Provider 32 -0.001
Provider 33 0.002
0.6
1
1.4
1.8
2.2
Actual
Bootstrap
90% Sample


Total Cost Index
Variance
Provider 34 0.000
Provider 35 0.000
Provider 36 0.000
Provider 37 -0.003
Provider 38 0.000
Provider 39 0.003
Provider 40 0.000
Provider 41 -0.001
Provider 42 0.001
Provider 43 0.001
Provider 44 0.003
Provider 45 0.007
Provider 46 0.001
Provider 47 0.002
Provider 48 0.000
Provider 49 0.000
Provider 50 -0.002
Provider 51 0.000
Provider 52 0.008
Provider 53 0.004
Provider 54 0.001
Provider 55 -0.003
Provider 56 0.000
Provider 57 0.001
Provider 58 0.005
Provider 59 -0.001
Provider 60 -0.001
Provider 61 0.005
Provider 62 -0.004
Provider 63 -0.005
Provider 64 -0.002
Provider 65 0.002
Provider 66 -0.006
0.6
1
1.4
1.8
2.2
Actual
Bootstrap
90% Sample


Resource Use Index
Variance
Provider 34 0.000
Provider 35 0.000
Provider 36 0.000
Provider 37 -0.003
Provider 38 0.000
Provider 39 0.002
Provider 40 0.000
Provider 41 0.000
Provider 42 0.001
Provider 43 0.000
Provider 44 0.002
Provider 45 0.006
Provider 46 0.000
Provider 47 0.002
Provider 48 0.000
Provider 49 0.002
Provider 50 -0.002
Provider 51 -0.002
Provider 52 0.005
Provider 53 0.004
Provider 54 0.000
Provider 55 -0.003
Provider 56 0.000
Provider 57 0.001
Provider 58 0.004
Provider 59 -0.001
Provider 60 0.000
Provider 61 0.003
Provider 62 -0.003
Provider 63 -0.004
Provider 64 -0.001
Provider 65 0.000
Provider 66 -0.002
0.6
1
1.4
1.8
2.2
Actual
Bootstrap
90% Sample

| | Total Cost Index | | | | | Resource Use Index | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Provider Group | 90% Sample | Bootstrap | Actual | Variation between Actual and Bootstrap | Variation between Actual and 90% | 90% Sample | Bootstrap | Actual | Variation between Actual and Bootstrap | Variation between Actual and 90% |
| Provider 01 | 0.836 | 0.836 | 0.836 | (0.001) | (0.000) | 0.930 | 0.930 | 0.931 | (0.000) | (0.000) |
| Provider 02 | 0.841 | 0.845 | 0.842 | 0.004 | (0.001) | 0.951 | 0.955 | 0.951 | 0.003 | (0.001) |
| Provider 03 | 0.849 | 0.848 | 0.849 | (0.001) | (0.000) | 0.914 | 0.913 | 0.915 | (0.001) | (0.000) |
| Provider 04 | 0.868 | 0.869 | 0.868 | 0.001 | (0.000) | 0.968 | 0.970 | 0.968 | 0.002 | (0.000) |
| Provider 05 | 0.873 | 0.872 | 0.873 | (0.002) | (0.001) | 0.825 | 0.823 | 0.826 | (0.002) | (0.001) |
| Provider 06 | 0.883 | 0.885 | 0.884 | 0.001 | (0.001) | 0.960 | 0.963 | 0.961 | 0.002 | (0.001) |
| Provider 07 | 0.892 | 0.895 | 0.891 | 0.004 | 0.001 | 0.969 | 0.971 | 0.968 | 0.003 | 0.001 |
| Provider 08 | 0.902 | 0.903 | 0.903 | 0.000 | (0.000) | 0.940 | 0.941 | 0.940 | 0.000 | (0.000) |
| Provider 09 | 0.903 | 0.902 | 0.903 | (0.000) | 0.000 | 0.992 | 0.991 | 0.992 | (0.001) | 0.000 |
| Provider 10 | 0.904 | 0.904 | 0.904 | (0.000) | (0.000) | 0.981 | 0.981 | 0.981 | (0.000) | (0.000) |
| Provider 11 | 0.910 | 0.911 | 0.910 | 0.001 | 0.000 | 0.999 | 1.001 | 0.999 | 0.002 | 0.001 |
| Provider 12 | 0.911 | 0.911 | 0.911 | (0.000) | (0.000) | 0.980 | 0.980 | 0.980 | (0.000) | (0.000) |
| Provider 13 | 0.917 | 0.916 | 0.917 | (0.001) | (0.000) | 0.988 | 0.988 | 0.987 | 0.000 | 0.000 |
| Provider 14 | 0.918 | 0.918 | 0.917 | 0.000 | 0.000 | 0.947 | 0.947 | 0.946 | 0.001 | 0.000 |
| Provider 15 | 0.918 | 0.917 | 0.918 | (0.001) | (0.000) | 0.922 | 0.921 | 0.922 | (0.001) | (0.000) |
| Provider 16 | 0.926 | 0.926 | 0.926 | 0.000 | 0.000 | 1.019 | 1.020 | 1.019 | 0.000 | 0.000 |
| Provider 17 | 0.926 | 0.928 | 0.926 | 0.002 | (0.000) | 0.973 | 0.974 | 0.973 | 0.002 | (0.000) |
| Provider 18 | 0.945 | 0.943 | 0.944 | (0.001) | 0.000 | 0.894 | 0.892 | 0.893 | (0.001) | 0.000 |
| Provider 19 | 0.945 | 0.945 | 0.945 | (0.000) | (0.000) | 1.006 | 1.006 | 1.007 | (0.000) | (0.000) |
| Provider 20 | 0.957 | 0.957 | 0.958 | (0.000) | (0.000) | 0.981 | 0.981 | 0.981 | (0.000) | (0.000) |
| Provider 21 | 0.959 | 0.960 | 0.959 | 0.001 | 0.000 | 1.012 | 1.012 | 1.011 | 0.001 | 0.000 |
| Provider 22 | 0.960 | 0.962 | 0.960 | 0.001 | (0.000) | 0.869 | 0.871 | 0.870 | 0.001 | (0.001) |
| Provider 23 | 0.962 | 0.964 | 0.963 | 0.001 | (0.001) | 1.009 | 1.012 | 1.011 | 0.001 | (0.002) |
| Provider 24 | 0.974 | 0.973 | 0.973 | (0.001) | 0.000 | 1.033 | 1.032 | 1.032 | (0.001) | 0.000 |
| Provider 25 | 0.975 | 0.974 | 0.974 | (0.000) | 0.001 | 0.987 | 0.986 | 0.986 | 0.001 | 0.001 |
| Provider 26 | 0.976 | 0.976 | 0.976 | (0.000) | (0.000) | 0.997 | 0.997 | 0.997 | (0.000) | (0.000) |
| Provider 27 | 0.979 | 0.978 | 0.978 | (0.000) | 0.000 | 1.120 | 1.120 | 1.119 | 0.000 | 0.000 |
| Provider 28 | 0.985 | 0.985 | 0.985 | (0.000) | (0.000) | 1.041 | 1.040 | 1.041 | (0.000) | (0.000) |
| Provider 29 | 1.007 | 1.010 | 1.008 | 0.002 | (0.000) | 0.939 | 0.941 | 0.939 | 0.002 | (0.000) |
| Provider 30 | 1.014 | 1.013 | 1.013 | (0.000) | 0.001 | 1.024 | 1.022 | 1.022 | 0.000 | 0.002 |
| Provider 31 | 1.013 | 1.014 | 1.013 | 0.001 | (0.000) | 0.910 | 0.911 | 0.910 | 0.000 | 0.000 |
| Provider 32 | 1.019 | 1.016 | 1.019 | (0.003) | (0.000) | 1.042 | 1.040 | 1.042 | (0.001) | 0.000 |
| Provider 33 | 1.022 | 1.026 | 1.023 | 0.003 | (0.001) | 1.141 | 1.144 | 1.142 | 0.002 | (0.001) |
| Provider 34 | 1.026 | 1.026 | 1.026 | 0.000 | (0.000) | 1.017 | 1.017 | 1.017 | 0.000 | (0.000) |
| Provider 35 | 1.028 | 1.028 | 1.028 | (0.000) | (0.000) | 0.961 | 0.961 | 0.961 | 0.000 | (0.000) |
| Provider 36 | 1.038 | 1.037 | 1.037 | (0.000) | 0.000 | 1.140 | 1.139 | 1.139 | 0.000 | 0.000 |
| Provider 37 | 1.040 | 1.037 | 1.040 | (0.003) | (0.000) | 1.171 | 1.168 | 1.171 | (0.003) | (0.000) |
| Provider 38 | 1.052 | 1.051 | 1.051 | 0.000 | 0.000 | 0.968 | 0.967 | 0.968 | (0.000) | 0.000 |
| Provider 39 | 1.066 | 1.069 | 1.066 | 0.003 | 0.000 | 0.907 | 0.908 | 0.906 | 0.002 | 0.001 |
| Provider 40 | 1.066 | 1.066 | 1.066 | (0.000) | (0.000) | 1.116 | 1.116 | 1.116 | 0.000 | 0.000 |
| Provider 41 | 1.070 | 1.070 | 1.071 | (0.001) | (0.000) | 1.124 | 1.124 | 1.124 | (0.000) | (0.000) |
| Provider 42 | 1.074 | 1.075 | 1.074 | 0.001 | (0.001) | 0.978 | 0.979 | 0.979 | 0.001 | (0.000) |
| Provider 43 | 1.083 | 1.084 | 1.084 | 0.001 | (0.001) | 0.915 | 0.917 | 0.917 | 0.000 | (0.001) |
| Provider 44 | 1.100 | 1.104 | 1.101 | 0.003 | (0.001) | 1.020 | 1.023 | 1.021 | 0.002 | (0.001) |
| Provider 45 | 1.107 | 1.114 | 1.107 | 0.007 | 0.000 | 0.888 | 0.895 | 0.888 | 0.006 | (0.000) |
| Provider 46 | 1.112 | 1.113 | 1.112 | 0.001 | (0.000) | 0.916 | 0.916 | 0.916 | 0.000 | 0.000 |
| Provider 47 | 1.113 | 1.114 | 1.113 | 0.002 | 0.000 | 0.908 | 0.910 | 0.908 | 0.002 | 0.000 |
| Provider 48 | 1.117 | 1.118 | 1.118 | (0.000) | (0.001) | 1.022 | 1.023 | 1.022 | 0.000 | (0.001) |
| Provider 49 | 1.171 | 1.171 | 1.171 | 0.000 | (0.000) | 0.961 | 0.964 | 0.962 | 0.002 | (0.001) |
| Provider 50 | 1.180 | 1.179 | 1.182 | (0.002) | (0.002) | 1.081 | 1.080 | 1.082 | (0.002) | (0.001) |
| Provider 51 | 1.187 | 1.188 | 1.188 | (0.000) | (0.001) | 1.050 | 1.049 | 1.051 | (0.002) | (0.000) |
| Provider 52 | 1.191 | 1.199 | 1.191 | 0.008 | 0.000 | 0.899 | 0.904 | 0.899 | 0.005 | 0.000 |
| Provider 53 | 1.201 | 1.205 | 1.201 | 0.004 | (0.000) | 0.968 | 0.972 | 0.968 | 0.004 | (0.000) |
| Provider 54 | 1.203 | 1.203 | 1.203 | 0.001 | 0.001 | 0.927 | 0.927 | 0.926 | 0.000 | 0.001 |
| Provider 55 | 1.254 | 1.251 | 1.253 | (0.003) | 0.001 | 0.990 | 0.987 | 0.990 | (0.003) | 0.001 |
| Provider 56 | 1.255 | 1.255 | 1.255 | (0.000) | 0.000 | 1.028 | 1.027 | 1.028 | (0.000) | 0.000 |
| Provider 57 | 1.256 | 1.259 | 1.258 | 0.001 | (0.001) | 0.941 | 0.944 | 0.942 | 0.001 | (0.001) |
| Provider 58 | 1.266 | 1.274 | 1.268 | 0.005 | (0.002) | 1.123 | 1.129 | 1.125 | 0.004 | (0.002) |
| Provider 59 | 1.294 | 1.292 | 1.293 | (0.001) | 0.000 | 1.051 | 1.050 | 1.051 | (0.001) | 0.000 |
| Provider 60 | 1.359 | 1.359 | 1.359 | (0.001) | (0.000) | 0.940 | 0.940 | 0.940 | (0.000) | (0.000) |
| Provider 61 | 1.359 | 1.365 | 1.359 | 0.005 | (0.001) | 1.073 | 1.076 | 1.073 | 0.003 | (0.001) |
| Provider 62 | 1.423 | 1.420 | 1.424 | (0.004) | (0.001) | 0.958 | 0.956 | 0.959 | (0.003) | (0.001) |
| Provider 63 | 1.472 | 1.467 | 1.472 | (0.005) | (0.000) | 0.988 | 0.984 | 0.988 | (0.004) | 0.000 |
| Provider 64 | 1.538 | 1.535 | 1.538 | (0.002) | 0.000 | 0.965 | 0.964 | 0.965 | (0.001) | 0.000 |
| Provider 65 | 1.669 | 1.674 | 1.672 | 0.002 | (0.002) | 1.056 | 1.057 | 1.057 | (0.000) | (0.002) |
| Provider 66 | 2.027 | 2.022 | 2.028 | (0.006) | (0.000) | 1.398 | 1.396 | 1.399 | (0.002) | (0.000) |

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The results of the Bootstrap and Random Sample tests allow us to confidently conclude that the measures will reliably decipher RUI performance between levels of analysis (e.g. provider group).

- The bootstrap results indicate that the RUIs are reliable as the provider variation within all groups is <1% whereas the variation between groups spans >110%.

Reliability Paper describes the provider group results of testing in detail:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188105.pdf

---

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)

x☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

x☐ **Performance measure score**

x☐ **Empirical validity testing**

x☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

*Prior submission: Please see Appendix A (page 14) for validity testing results from prior submission. The method of testing used for current resubmission is the same methodology used in prior submission.*

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

A Validity Analysis was performed on the HealthPartners' Total Resource Use measure which indicates the results accurately reflect the performance levels of provider groups. When used in conjunction with the Total Cost of Care measure, the measure also accurately identifies the price (per unit cost) performance levels of providers.

Detailed testing can be found in the Validity paper:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

Critical data elements

Non-risk adjusted correlations between ACG and Total Resource Use, Total Cost Relative Resource Values (resource use) and utilization metrics were calculated.

Performance Measure Score

Risk adjusted Resource Use Index correlations to known risk adjusted utilization metrics were calculated.

Empirical testing of validity and overview of face validity policy and procedure

An assessment of high and low performing provider groups supports the relationship between risk adjusted utilization metrics and Resource Use Index.

The face validity process is conducted by transparently sharing results and methods with provider groups measured and allowing a 45-day comment period prior to public display of provider group results.

HealthPartners has a Policy and Procedure Review Process and executes it annually with each release of provider groups' performance and measurement results. Disclosure to providers includes:

1. Transparent reporting of measurement methodology
2. Providing comparative performance results with information on statistical reliability to providers

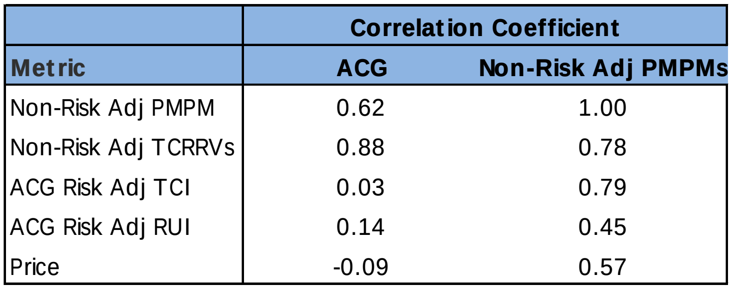
3. Providing an explanation of the results at least 45 days prior to their use in public reporting or business applications
4. Notifying providers of how the information will be used
5. A process by which providers can notify the plan of additional information or corrections

Public reporting of provider group measurement results:
https://www.healthpartners.com/public/cost-and-quality/index.html
Publicly available methods of rate calculations for transparency:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_033165.pdf

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

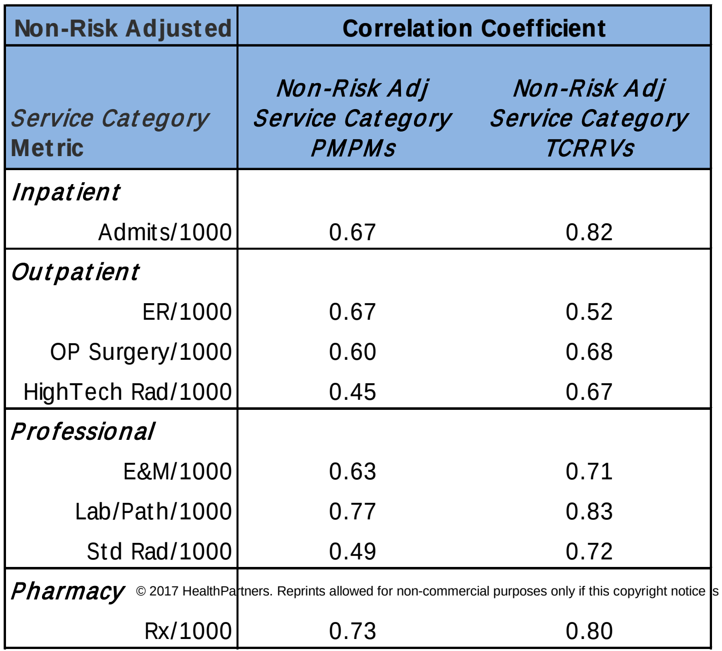
The correlation coefficients are included below for testing validity of the measure components and validity of the Total Resource Use measure. Interpretation accompanies the tables of results below to provide context. However, please reference the paper to follow the complete analytical pathway with context and reasoning to conclude the measure is valid.
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

**<u>Validity of Measure Components</u>**
**Correlations Between ACG Score, Non-Risk Adjusted Per Member Per Month (PMPMs), Non-Risk Adjusted Total Cost Relative Resource Values (TCRRVs), and Risk Adjusted RUI**



| | Correlation Coefficient | |
|---|---|---|
| Metric | ACG | Non-Risk Adj PMPMs |
| Non-Risk Adj PMPM | 0.62 | 1.00 |
| Non-Risk Adj TCRRVs | 0.88 | 0.78 |
| ACG Risk Adj TCI | 0.03 | 0.79 |
| ACG Risk Adj RUI | 0.14 | 0.45 |
| Price | -0.09 | 0.57 |

- There is a high correlation between ACG score and the non-risk adjusted PMPM and TCRRVs which indicates that the non-risk adjusted PMPM and the non-risk adjusted TCRRVs are a good measure of resource use.
- There is a low correlation between ACG score and the risk adjusted RUI. This indicates that the risk score of a provider has no impact on a provider's ability to be a high performer.
- There is a low correlation between price and ACG because ACGs measure expected resource use whereas price is not affected by the number or intensity of services received.



| Non-Risk Adjusted | Correlation Coefficient | |
|---|---|---|
| *Service Category* Metric | *Non-Risk Adj Service Category PMPMs* | *Non-Risk Adj Service Category TCRRVs* |
| *Inpatient* | | |
| Admits/1000 | 0.67 | 0.82 |
| *Outpatient* | | |
| ER/1000 | 0.67 | 0.52 |
| OP Surgery/1000 | 0.60 | 0.68 |
| HighTech Rad/1000 | 0.45 | 0.67 |
| *Professional* | | |
| E&M/1000 | 0.63 | 0.71 |
| Lab/Path/1000 | 0.77 | 0.83 |
| Std Rad/1000 | 0.49 | 0.72 |
| *Pharmacy* | | |
| Rx/1000 | 0.73 | 0.80 |

**Correlations Between the Non-Risk Adjusted Place of Service Metrics and Non-Risk Adjusted PMPMs & Non-Risk Adjusted TCRRVs**

**Inpatient:** There should be and are strong correlations between the admit rate to the non-risk adjusted PMPMs and non-risk adjusted TCRRVs as the only two factors not measured by the admits are the intensity and unit cost of the services performed.

**Outpatient:** There should be and are moderate correlations between the ER, outpatient surgery, and high tech radiology rates to the non-risk adjusted PMPMs and non-risk adjusted TCRRVs as these three utilization metrics combine to encompass approximately 65% of the total outpatient spend.

**Professional:** There should be and are moderate correlations between the E&M visits, Lab/Path services, and standard radiology to the non-risk adjusted PMPMs and non-risk adjusted TCRRVs as they represent 45% of the professional spend, but are also good indicators of patients that consume medical services.

**Pharmacy:** There should be strong correlations between the pharmacy prescribing rates to the non-risk adjusted PMPMs and non-risk adjusted TCRRVs as the only factor that is not accounted for in the Rx prescribing rate metric is the intensity of the drug prescribed. The intensity includes generic usage as well as the variation in cost between drugs.

Since the ACG score, non-risk adjusted PMPMs and non-risk adjusted TCRRVs are a measure of the consumption of health care services, there should be and are strong correlations between these values and known utilization metrics.

Composite Utilization: A utilization metric was created by weighting each of the underlying utilization metrics by the place of service percent of resources it represents of the total resources by each provider group.

Composite Utilization Metric by Provider Group =
Inpatient (Admit Rate x Inpatient Resource Use %) +
Outpatient (Average (ER rate, OP Surg Rate, High Tech Rad Rate) x Outpatient Resource Use %) +
Professional (Average (E&M rate, Lab/Path Rate, Std Rad) x Professional Resource Use %) +
Pharmacy (Rx rate x Pharmacy Resource Use %)

| Non-Risk Adjusted | Correlation Coefficient | | |
|---|---|---|---|
| Metric | ACG | Non-Risk Adj PMPMs | Non-Risk Adj TCRRVs |
| Composite Utilization | 0.74 | 0.69 | 0.87 |

The non-risk adjusted resource composite is highly correlated with ACGs, non-risk adjusted PMPMs and non-risk adjusted TCRRVs.

## Validity of Total Resource Use Measure

### Correlations Between the Risk Adjusted Place of Service Metrics and TCI and RUI

- Total Resource Use is correlated with TCI as expected.
- Professional RUI is highly correlated with overall RUI, supporting the notion primary care providers are integral in the management of total costs and resources.
- Hospital-based RUI has a lower correlation than professional as a lower proportion of patients require hospital based care.

| Risk Adjusted | Correlation Coefficient | | |
|---|---|---|---|
| Metric | TCI | RUI | Price |
| Hospital TCI | 0.74 | | |
| Prof TCI | 0.73 | | |
| Rx TCI | 0.16 | | |
| Hospital RUI | | 0.30 | |
| Prof RUI | | 0.74 | |
| Total RUI | 0.39 | | |
| Hospital Price | | | 0.86 |
| Prof Price | | | 0.83 |
| Total Price | 0.87 | | |

### Correlations Between Risk Adjusted Place of Service Utilization Metrics and Corresponding RUI

| Risk Adjusted | Correlation Coefficient | |
|---|---|---|
| *Service Category* **Metric** | *Service Category TCIs* | *Service Category RUIs* |
| ***Inpatient*** | | |
| Admit Rate | 0.78 | 0.82 |
| ***Outpatient*** | | |
| ER Cnt | 0.68 | 0.46 |
| OP Surgery | 0.55 | 0.49 |
| High Tech Rad | 0.21 | 0.37 |
| ***Professional*** | | |
| E&M Visits | 0.48 | 0.70 |
| Lab/Path | 0.59 | 0.54 |
| Std Rad | 0.48 | 0.38 |
| ***Pharmacy*** | | |
| Rx Count | 0.25 | |

**Inpatient:** There is a high correlation between the risk adjusted admit rate and the inpatient RUI. This would indicate that the higher the risk adjusted admit rate the more likely a provider will have a higher than average RUI.

**Outpatient:** There is a moderate correlation between the risk adjusted ER count and the outpatient RUI. This would indicate that the higher the risk adjusted ER counts the more likely a provider will have a higher than average outpatient RUI.

High tech radiology having less of a correlation to the outpatient RUI is an indication that these services are not the driving force behind the outpatient RUI performance as they are not as prevalent.

**Professional:** The professional utilization metrics are moderately correlated to the professional RUI.

This result is as expected as the professional place of service includes a significant amount of services beyond these three utilization measures (other professional services = 55%).

It is also as expected because having higher than average utilization on diagnostic or management based services does not necessarily indicate a higher resource consuming patient.

| Risk Adjusted | Correlation Coefficient | Correlation Coefficient |
|---|---|---|
| **Metric** | **TCI** | **RUI** |
| Composite Utilization | 0.72 | 0.52 |

The indexed Total Resource Use measure has a high correlation to a risk adjusted composite utilization index, which was developed as a proxy to measure total resource consumption.

*Prior submission: Please see Appendix A (page 14) for prior submission results.*

In addition, the Total Resource Use measure was analyzed over time (2013 through 2015) to demonstrate stability and sensitivity to provider changes or improvement initiatives. Providers' performance across all three measures is relatively consistent across all three years and results are shown in the table below. The factors that drive variation between years within a provider are cost per unit and resource use management.

The results show that TCI has the most variation as it combines the changes for both price and resource use. The results also show that there is more variation in resource use over time than price. This indicates that providers are receiving similar price increases, but how providers are managing their patients' resource use is contributing more to the variation seen in costs.

| | TCI | | | | Price | | | | RUI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Provider Group Size | 25th Percentile | Average | Median | 75th Percentile | 25th Percentile | Average | Median | 75th Percentile | 25th Percentile | Average | Median | 75th Percentile |
| <1,000 | 0.04 | 0.07 | 0.07 | 0.11 | 0.02 | 0.04 | 0.03 | 0.05 | 0.03 | 0.05 | 0.05 | 0.09 |
| 1,000-2,000 | 0.03 | 0.08 | 0.07 | 0.11 | 0.02 | 0.03 | 0.03 | 0.04 | 0.03 | 0.06 | 0.07 | 0.09 |
| 2,000+ | 0.01 | 0.03 | 0.03 | 0.04 | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 | 0.03 | 0.03 | 0.05 |

*Prior submission: Please see Appendix A (page 12) for prior submission results.*

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The Total Resource Use measure is valid as the critical data elements and the criteria applied produce a measure that accurately assesses various levels of performance. The norms in the measure are the network averages from the healthcare information derived from the MN market from included entities.

The Validity paper describes the results and conclusions from testing in detail: https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

In summary, the Total Resource Use measure accurately and consistently identified providers that are low or high performers with conclusions supported by known utilization measures.

There are high correlations between non-risk adjusted PMPM, ACG score and non-risk adjusted TCRRVs which indicate they are good measures of resources.

The ACGs, non-risk adjusted PMPMs, and non-risk adjusted TCRRVs have similar correlations to all utilization metrics which indicates the TCRRVs are performing as expected and are a solid measure of resources.

The indexed Resource Use measure scores have a high correlation (0.52) to a risk adjusted composite utilization index score, which was developed as a proxy to measure total resource consumption.

The Total Resource Use measure differentiates between provider groups accurately as supported by the risk adjusted service utilization metrics.

## 2b3. EXCLUSIONS ANALYSIS

**NA □ no exclusions — *skip to section 2b4***

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps***—***do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

The HealthPartners' Total Resource Use measure is a full population-based measure, with members under age 1, members 65+ and members with less than 9 months of enrollment excluded to ensure an accurate risk assessment is made on the population.

- Members over age 64
- Members under age 1
- Member enrollment less than nine months during the one year measurement time window
- TCRRVs per member up to 125,000 are included; TCRRVs per member above 125,000 are excluded (truncated)

*Prior submission: For this maintenance submission, the only change to HealthPartners Total Resource Use measure from prior submission is the truncation level. The total TCRRV truncation level for a member's combined medical and pharmacy claims has increased from 100,000 to 125,000 TCRRVs to account for the natural rise in healthcare costs over the past several years.*

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Results from testing truncation level at 125,000 TCRRVs can be found in the Validity paper. No other changes to measure criteria have occurred since endorsement.
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

When the Resource Use measure is deployed independently of the TCOC measure the 125,000 TCRRV truncation level, which is the same truncation level as TCOC, can be applied because the TCRRVs are calibrated to reflect a standardized total paid. The TCOC truncation was increased from $100,000 to $125,000. Given medical inflation has been 2-4% per year recently, it is necessary to increase the spend truncation to account for the natural rise in healthcare costs.

Since the model needs to remain stable year over year, the truncation level also needs to remain stable, with only periodic updates. The $125,000 truncation level returns the model to its original NQF endorsed state in terms of R-squared, percent of dollars included in the model.

The truncation level for Resource Use when used in conjunction with TCOC is variable by member as the Total Care Relative Resource Use Values (TCRRVs) are truncated in the same proportion as the total paid amount. The practical effect is the price (i.e., total paid amount/TCRRV) for the services for the truncated members remains constant as the total paid is reduced using the same factor as the TCRRVs.

The following exclusions and decision points remain unchanged from the original endorsed measures.

Nine month continuous enrollment – A nine month continuous enrollment was selected to balance business operations. Nine months allows for partial year enrollee. There was very little statistical difference in R-squared between six and twelve months.

Infants, under age one are excluded due to slightly higher R-squared of the population without newborns, the required nine months enrollment criteria and variability in newborn costs, newborns under age one were excluded from the total resource use measure.
Members over age 64 due are excluded due to potential incomplete claims data of Medicare eligible beneficiary.

| Resource Use Measure Population Exclusion Funnel | Percent of Members | Percent of Total Paid |
|---|---|---|
| All Commercial Members | 100% | 100% |
| Members over 1 | 99% | 98% |
| Members between 1-64 | 96% | 91% |
| Members age 1-64 and enrolled 9 months | 78.3% | 84% |
| Truncated at 125,000* | 0.28% | 79.2% |
| Member and Spend included | 78.3% | 79.2% |

** Members are not removed from the measures*

## 2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.***

**2b4.1. What method of controlling for differences in case mix is used?**

☐ **No risk adjustment or stratification**
☐ **Statistical risk model with 0 risk factors**
☐ **Stratification by 0 risk categories**
x☐ **Other,** Johns Hopkins ACG System on commercially covered population

**2b4.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

The Total Resource Use measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). The measure is also limited by insurance coverage to commercial only.

The ACG System is a statistically valid and broadly adopted risk grouper in both academic and non-academic settings with methodology derived from diagnosis information. Information about the development of the grouper can be found here: http://acg.jhsph.org/; additionally please refer to the ACG Technical Reference Guide for supporting material: https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf

ACG Grouper:

- Adjusted Clinical Groups (ACG System) were developed by Johns Hopkins University and allow comparisons between populations with varying illness burdens based on diagnoses, age and gender.
- Each unique member is assigned one of 93 ACG actuarial cells, which has a corresponding weight that reflects relative illness burden (e.g. relative expected resource consumption). Attributed members are assigned a risk score based on diagnoses on claims from the performance measurement period, as well as member age and gender

ACG-cell Risk Weights/Coefficients:

- The ACG risk weights measure relative resource variation between ACG actuarial cells/codes. Please see page 30-34 of the reference guide to view each ACG-cell risk weight. https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf
- Multiply each member's ACG weight by their eligible member months creating a total member ACG weight.

ACG Score:

- Each provider's attributed member ACG weights are summed to the provider level and divided by the sum of the attributed member months creating an ACG score for the provider.
- The provider's average ACG score is indexed to all attributed member's plan average ACG score.
- A member's total member ACG weight is updated to correspond with each year the Total Resource Use measure is measured.

Each of the 93 ACG actuarial cells can be considered a covariate of the multivariate risk model with the cell weights being the coefficients. The ACG cells are non-linear composites of the three risk factors: age, gender, diagnosis. Each member is assigned one of 93 covariates in the multivariate model and is based on the member's combination of age, gender and complete history of diagnosis codes.

**2b4.2. If an outcome or resource use component measure is <u>not risk adjusted or stratified</u>, provide <u>rationale and analyses</u> to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

Not applicable. All measures are clinically risk adjusted and limited to the commercial population.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

The Total Resource Use measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). The measure is also limited by insurance coverage to the commercial population.

The ACG System is a statistically valid and broadly adopted risk grouper in both academic and non-academic settings with methodology derived from diagnosis information. Information about the development of the grouper can be found here: http://acg.jhsph.org/; additionally please refer to the ACG Technical Reference Guide for supporting material:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf

The ACG System assigns International Classification of Disease (ICD) diagnosis codes to 32 diagnosis groups – Aggregated Diagnosis Groups (ADGs). The assignment method is included in the ACG software for all codes. Diagnosis codes mapped to a given ADG are clinically similar and have similar expected need for healthcare resources. The assignment criteria is based on features of a condition that help predict duration and intensity of resource use. Five clinical criteria are used to determine assignment of codes: duration, severity, diagnostic certainty, type of etiology, and expected need for specialty care. The 32 ADGs are listed on pages 4-6 in the reference guide, along with a table on pages 8-10 that provides guidance on how the five criteria are applied to each ADG.

Adjusted Clinical Group actuarial cells (ACGs) build off of the ADG assignment logic described and are used to determine the morbidity profile of patient populations to more fairly assess provider performance and allow for equitable comparisons of utilization and outcomes. ACGs are defined by morbidity, age, and sex and are person-focused to categorize patients' illnesses. Based on the pattern of morbidities, the ACG approach assigns each individual to a single ACG category. The ACG assignment process can be found on page 12 of the reference guide.

After applying measure criteria, which includes limitation to commercial only and clinical risk adjustment, socioeconomic testing was conducted that considered income and education status as potential factors beyond those already adjusted for. Methodology and testing results can be found here:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

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

The risk factors included in ACG risk grouper were determined in the development of Johns Hopkins ACG risk grouper and are not available to the general public. The performance of the risk groupers are the basis for verifying the risk factors included in the model are sufficient to address clinic risk variation. The Society of Actuaries Accuracy of Claims-Based Risk Scoring Models (2016) findings also indicate the reliability and validity of the ACG risk grouper.
https://www.soa.org/Files/Research/research-2016-accuracy-claims-based-risk-scoring-models.pdf

**2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)**

After risk adjusting for age, gender, and clinical risk, and limiting by insurance type, income does not significantly impact a patient's total resource use. As a potential practical use case example, the study also evaluated Resource Use provider group performance and found there was no discernible difference in performance when adjusting for income. The provider group analysis focused on the Resource Use measure to remove any bias based on price. The study considered two different data sources to study income variation, Census tract data and a commercially licensed data source available to HealthPartners with more specific income data.

https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

The study utilized two independent data sources to evaluate income. The first was U.S. Census tracts. As defined by the U.S. Department of Commerce, "Census tracts are small, relatively permanent statistical subdivisions of a county or equivalent entity that are updated by local participants prior to each decennial census as part of the Census Bureau's Participant Statistical Areas Program. The Census Bureau delineates census tracts in situations where no local participant existed or where state, local, or tribal governments declined to participate. The primary purpose of census tracts is to provide a stable set of geographic units for the presentation of statistical data.

Census tracts generally have a population size between 1,200 and 8,000 people, with an optimum size of 4,000 people. A census tract usually covers a contiguous area; however, the spatial size of census tracts varies widely depending on the density of settlement. Census tract boundaries are delineated with the intention of being maintained over a long time so that statistical comparisons can be made from census to census. Census tracts occasionally are split due to population growth or merged as a result of substantial population decline."[5] As noted, tracts estimate income by a general area and are not highly specific, introducing potential error and bias in the model.

HealthPartners utilized an additional data source to more accurately assess household income for purposes of this study. HealthPartners commercially licenses and has access to a large consumer database for other business purposes which gave us the ability to evaluate income with more specificity at the household level. Recognizing that it may not be feasible for all users to access a commercial database, HealthPartners pursued this deeper evaluation to more broadly understand the important question of whether or not to adjust resource use performance measures by socioeconomic status independent of data availability. Household level income is derived using the midpoint of defined ranges of income by household (e.g. $20,000-$30,000) and capped at $250,000. Using the midpoint of a range introduces potential error in the evaluation whereas self-reported individual or household income would be most accurate.

## Population-Based Evaluation

The evaluation tested the inclusion of income in addition to the factors already included in the measure specifications - age, gender, and clinical risk. Detailed measure criteria can be found in HealthPartners Technical Guidelines.

The study population included HealthPartners' full book of business of members, Commercial and Medicaid with TCOC criteria applied using services and claims generated throughout the 2015 time period. The study population included more than 530,000 members.

Three multiple linear regression models were created, each with one of the three metrics of interest as the dependent variable (total reimbursed amount per member per month, resource use per member per month, and price). Each model was identical in the use of income, ACG risk score, and insurance product (commercial vs Medicaid) as the independent variables. Resource use, reimbursed amount, price, and ACG scores were log transformed prior to developing the regression models to address the skewed nature of the data and adjust for heteroscedasticity. Insurance product was treated as a binary variable (commercial = 1, Medicaid = 0). The resulting coefficients were analyzed in terms of a 1% increase from average and their corresponding effect on the dependent variables.

Additionally, a model was created using only the endorsed measure criteria for the Resource Use measure (i.e. ACG and product only as the independent variables). The $R^2$ statistic from this model was compared against the $R^2$ statistic from the model that included income as an independent variable, allowing us to quantify the predictive value of income on resource use.

The same regression statistics and models were used with the second, more robust data source available to HealthPartners. This data contained more accurate income information which was specific to household rather than tract, with household income defined using the midpoint or median of the income ranges. The more robust data source was available for 65% of HealthPartners' book of business members for 2015 and in the same proportions of commercial to Medicaid as in the previous evaluation.

**Provider Group Performance Evaluation**

A second evaluation was performed to provide a potential practical example of adjusting the TCOC and resource use measures by income using the Census and commercially licensed data sources. Resource Use Index was evaluated to remove known price variations between providers. HealthPartners' resource use results for its primary care network of commercial attributed members were used to evaluate provider group performance when adjusting for income. Medicaid was excluded from this evaluation as it has already been determined that provider performance results should be segmented by product.

There were 66 provider groups who met the measure criteria and were included in the evaluation using the Census tract data. The Total Resource Use measure is endorsed at a reliability level of 600 patients. Because the commercially licensed data source had available data for 65% of HealthPartners' book of business, there were 11 provider groups that failed to meet the 600 minimum and were excluded from the evaluation.

The variation between the average incomes using the Census tract data or the commercially licensed data source for each provider group was compared to the network average to adjust the provider's resource use index. It should be noted that while the adjustment can be made, the results should not be considered valid or reliable given the limitations inherent in each data source as described previously.

The regression analysis generated parameters that were translated into results based upon average cost, resource use, income, and ACG scores.

**Table of Regression results using Census Tract Data**

| Model | 1% Income Increase | 1% ACG increase | Commercial vs. Medicaid Membership |
|---|---|---|---|
| Total Reimbursement | $ (0.13) | $ 4.22 | $ 133.28 |
| Resource Use | $ 0.16 | $ 4.34 | $ (75.24) |
| Price | $ (0.28) | $ 0.07 | $ 205.36 |

| MODEL | R_SQUARED |
|---|---|
| Resource Use Endorsed Measure | 0.5788 |
| Resource Use Endorsed Measure + Income | 0.5792 |

Using Census tract data, a 1% increase in income resulted in a $0.13 decrease in total reimbursement, a $0.16 increase in resource use, and $0.28 decrease in price. The results highlight how significantly more the ACG score (clinical risk adjustment) and insurance product impact both the cost and resource use measures. For frame of reference, on average for the Midwest market, the total spend for a member per month (PMPM) is $400. The results of the evaluation show that a 1% increase in risk score accounts for a $4.22 or roughly 1% increase in PMPM.

Product also contributed significantly with there being a $133 dollar difference in cost between commercial and Medicaid. The variation in resource use was much less, however, still significant with Medicaid covered members utilizing $75 more dollars of resources. The fact that Medicaid's cost per service is approximately half that of commercial rates drives the differences between the TCOC and Resource Use results. The $R^2$ results further emphasize that ACG score and insurance type are the main drivers of cost and resource use variation and income does not provide any additional predictive power.

### Table of Regression results using Commercially Licensed Data Source

| Model | 1% Income Increase | 1% ACG increase | Commercial vs. Medicaid Membership |
|---|---|---|---|
| Total Reimbursement | $ (0.00) | $ 4.56 | $ 139.80 |
| Resource Use | $ 0.05 | $ 4.66 | $ (81.26) |
| Price | $ (0.07) | $ 0.06 | $ 218.13 |

| MODEL | R_SQUARED |
|---|---|
| Resource Use Endorsed Measure | 0.57318 |
| Resource Use Endorsed Measure + Income | 0.57321 |

Using the commercially purchased data source, with income by household, a 1% increase in income resulted in no change for total reimbursement, $0.05 increase in resource use, and $0.07 decrease in price. This is telling, as when using a data source that is more specific, income is even less impactful on TCOC and resource use while ACG and product type show similar results.

### Results– Provider Group Performance Evaluation

Provider group performance of the Resource Use measure was evaluated to test the impact of income adjustment on the Resource Use measure. Provider group results for both data sources, Census tract and commercially licensed, are shown below using HealthPartners' commercial provider network. The Resource Use Index (RUI) is calculated using the endorsed measure criteria. The second RUI is calculated using the endorsed measure criteria *with income adjustment.*

The Census tract data evaluated 66 provider groups and the commercially licensed data source evaluated 55 provider groups. Because the population of patients used between the two data sources is different, Provider Group 01 in the Census tract chart is not the same as Provider Group 01 in the commercially licensed chart. Provider group numbers in the Census tract chart are numbered based on ascending Total Cost Index found in the appendix of the study paper. Provider groups for both charts are sorted in ascending order using the RUI.

On average there was less than a 1% change in performance for provider groups when income was introduced into the model for the Resource Use measure when using Census tract data. This impact was reduced on average to less than a 0.25% when using the commercially licensed data source with more specific income data. Considering the Resource Use measure identifies provider performance levels (indices) that span greater than 167% as identified below, the less than 1% adjustment was considered insignificant when comparing provider performance. Provider Group charts begin on the following page.

### Census Tract Data Source

| | |
|---|---|
| RUI Min | 0.82 |
| RUI Max | 1.39 |
| RUI Max/Min % Difference | 167% |
| Average % change with income adjustment | 0.64% |

### Commercially Licensed Data Source

| | |
|---|---|
| RUI Min | 0.83 |
| RUI Max | 1.39 |
| RUI Max/Min % Difference | 167% |
| Average % change with income adjustment | 0.19% |

## Provider Group Detailed Results – Census Data

| Provider Group | TCI | RUI (endorsed measure) | RUI (endorsed measure + income) | Average Income | RUI Change | Pct RUI Change |
|---|---|---|---|---|---|---|
| Provider 05 | 0.87 | 0.82 | 0.83 | $51,182.66 | 0.0097 | 1.18% |
| Provider 22 | 0.96 | 0.87 | 0.87 | $60,871.13 | 0.0051 | 0.59% |
| Provider 45 | 1.11 | 0.88 | 0.89 | $51,196.01 | 0.0097 | 1.10% |
| Provider 52 | 1.19 | 0.89 | 0.90 | $57,184.20 | 0.0069 | 0.77% |
| Provider 18 | 0.94 | 0.89 | 0.90 | $53,262.19 | 0.0087 | 0.98% |
| Provider 39 | 1.07 | 0.90 | 0.91 | $52,994.97 | 0.0089 | 0.98% |
| Provider 47 | 1.11 | 0.90 | 0.92 | $48,573.69 | 0.0110 | 1.21% |
| Provider 31 | 1.01 | 0.91 | 0.91 | $54,522.47 | 0.0081 | 0.90% |
| Provider 46 | 1.11 | 0.91 | 0.92 | $55,143.95 | 0.0079 | 0.86% |
| Provider 43 | 1.08 | 0.91 | 0.93 | $49,821.34 | 0.0104 | 1.13% |
| Provider 03 | 0.85 | 0.92 | 0.91 | $74,230.68 | -0.0012 | -0.13% |
| Provider 15 | 0.92 | 0.92 | 0.93 | $54,236.28 | 0.0083 | 0.90% |
| Provider 54 | 1.20 | 0.92 | 0.93 | $59,432.45 | 0.0058 | 0.63% |
| Provider 60 | 1.36 | 0.93 | 0.93 | $57,038.75 | 0.0070 | 0.75% |
| Provider 01 | 0.84 | 0.93 | 0.94 | $59,923.30 | 0.0056 | 0.60% |
| Provider 29 | 1.01 | 0.94 | 0.94 | $56,657.27 | 0.0071 | 0.76% |
| Provider 57 | 1.26 | 0.94 | 0.95 | $46,884.50 | 0.0117 | 1.25% |
| Provider 08 | 0.90 | 0.94 | 0.94 | $74,671.53 | -0.0014 | -0.14% |
| Provider 64 | 1.54 | 0.95 | 0.96 | $50,511.60 | 0.0100 | 1.06% |
| Provider 62 | 1.42 | 0.95 | 0.96 | $51,481.13 | 0.0096 | 1.01% |
| Provider 49 | 1.17 | 0.95 | 0.95 | $63,017.62 | 0.0041 | 0.44% |
| Provider 14 | 0.92 | 0.95 | 0.94 | $85,046.08 | -0.0063 | -0.66% |
| Provider 02 | 0.84 | 0.96 | 0.96 | $75,988.94 | -0.0020 | -0.21% |
| Provider 35 | 1.03 | 0.96 | 0.97 | $53,580.68 | 0.0086 | 0.89% |
| Provider 53 | 1.20 | 0.96 | 0.97 | $60,513.25 | 0.0053 | 0.55% |
| Provider 42 | 1.07 | 0.96 | 0.97 | $56,581.35 | 0.0072 | 0.74% |
| Provider 38 | 1.05 | 0.96 | 0.97 | $53,033.63 | 0.0088 | 0.92% |
| Provider 06 | 0.88 | 0.96 | 0.96 | $78,737.12 | -0.0033 | -0.34% |
| Provider 63 | 1.47 | 0.97 | 0.97 | $68,995.93 | 0.0013 | 0.14% |
| Provider 04 | 0.87 | 0.97 | 0.97 | $63,162.88 | 0.0041 | 0.42% |
| Provider 07 | 0.89 | 0.97 | 0.96 | $87,449.16 | -0.0074 | -0.76% |
| Provider 17 | 0.93 | 0.97 | 0.97 | $75,724.77 | -0.0019 | -0.19% |
| Provider 20 | 0.96 | 0.98 | 0.98 | $81,800.09 | -0.0047 | -0.48% |

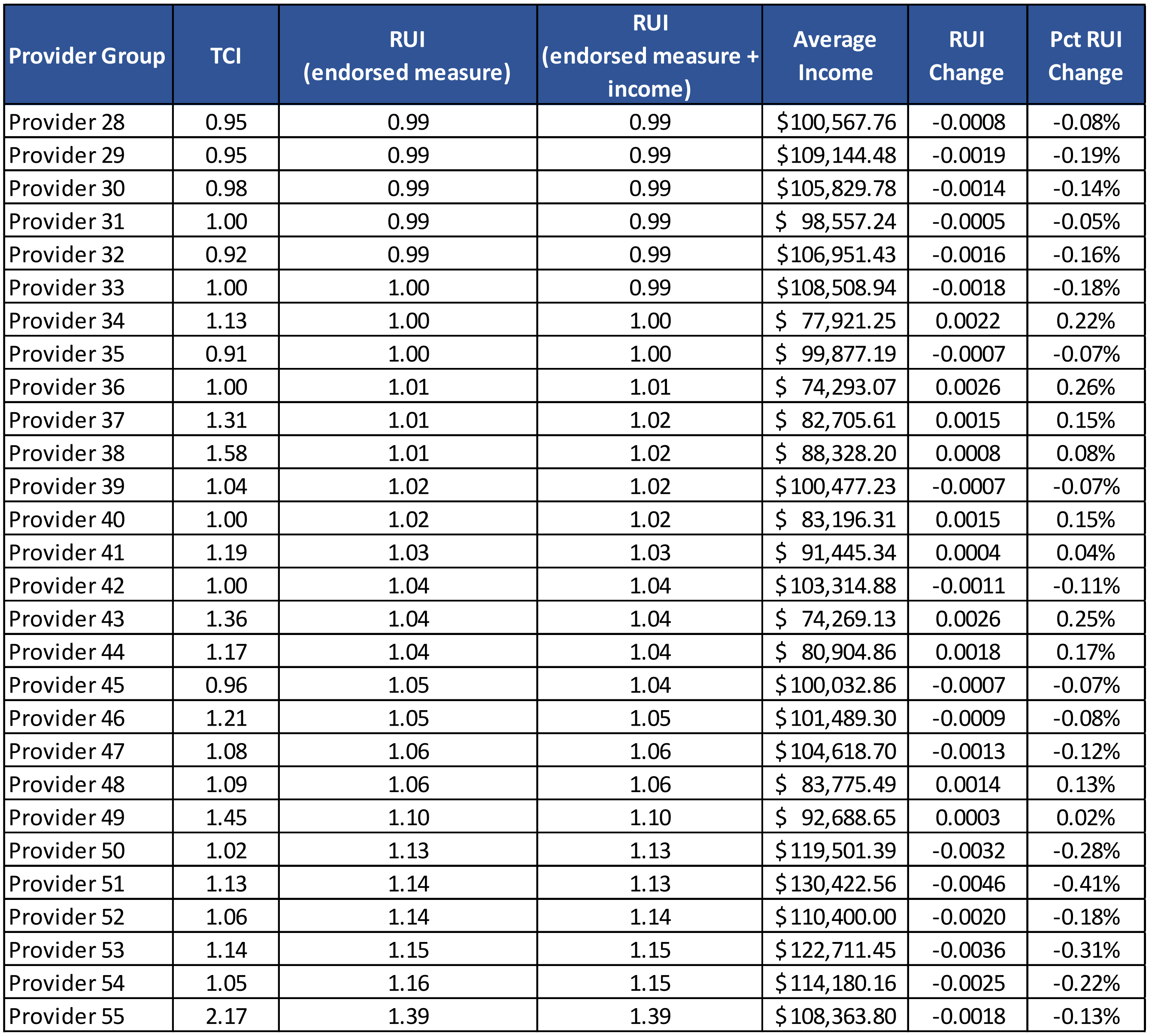
## Provider Group Detailed Results – Census Data - *continued*

| Provider Group | TCI | RUI<br>(endorsed measure) | RUI<br>(endorsed measure + income) | Average Income | RUI Change | Pct RUI Change |
|---|---|---|---|---|---|---|
| Provider 10 | 0.90 | 0.98 | 0.98 | $80,344.27 | -0.0040 | -0.41% |
| Provider 12 | 0.91 | 0.98 | 0.99 | $68,200.83 | 0.0017 | 0.17% |
| Provider 55 | 1.25 | 0.98 | 1.00 | $45,478.38 | 0.0124 | 1.26% |
| Provider 25 | 0.97 | 0.99 | 0.99 | $55,181.82 | 0.0078 | 0.79% |
| Provider 13 | 0.92 | 0.99 | 0.98 | $85,397.81 | -0.0064 | -0.65% |
| Provider 09 | 0.90 | 1.00 | 0.99 | $79,078.76 | -0.0034 | -0.35% |
| Provider 26 | 0.98 | 1.00 | 1.00 | $76,886.96 | -0.0024 | -0.24% |
| Provider 11 | 0.91 | 1.00 | 1.01 | $58,557.65 | 0.0062 | 0.62% |
| Provider 19 | 0.94 | 1.01 | 1.01 | $76,364.65 | -0.0022 | -0.21% |
| Provider 23 | 0.96 | 1.01 | 1.02 | $51,695.82 | 0.0095 | 0.94% |
| Provider 21 | 0.96 | 1.01 | 1.01 | $80,133.18 | -0.0039 | -0.39% |
| Provider 48 | 1.12 | 1.02 | 1.02 | $62,718.98 | 0.0043 | 0.42% |
| Provider 34 | 1.03 | 1.02 | 1.01 | $76,650.16 | -0.0023 | -0.23% |
| Provider 44 | 1.10 | 1.02 | 1.02 | $57,718.34 | 0.0066 | 0.65% |
| Provider 30 | 1.01 | 1.02 | 1.03 | $60,952.95 | 0.0051 | 0.50% |
| Provider 56 | 1.26 | 1.02 | 1.03 | $56,343.84 | 0.0073 | 0.71% |
| Provider 16 | 0.93 | 1.03 | 1.02 | $73,585.43 | -0.0009 | -0.08% |
| Provider 24 | 0.97 | 1.03 | 1.04 | $61,287.61 | 0.0050 | 0.48% |
| Provider 32 | 1.02 | 1.04 | 1.03 | $88,286.87 | -0.0078 | -0.75% |
| Provider 28 | 0.98 | 1.05 | 1.04 | $76,082.19 | -0.0020 | -0.19% |
| Provider 51 | 1.19 | 1.05 | 1.04 | $80,419.35 | -0.0041 | -0.39% |
| Provider 59 | 1.29 | 1.05 | 1.06 | $55,164.25 | 0.0078 | 0.75% |
| Provider 65 | 1.67 | 1.05 | 1.06 | $55,820.84 | 0.0075 | 0.72% |
| Provider 61 | 1.36 | 1.06 | 1.07 | $60,338.84 | 0.0054 | 0.51% |
| Provider 50 | 1.18 | 1.08 | 1.09 | $42,557.01 | 0.0138 | 1.28% |
| Provider 40 | 1.07 | 1.12 | 1.11 | $94,343.76 | -0.0106 | -0.95% |
| Provider 58 | 1.27 | 1.12 | 1.13 | $52,722.55 | 0.0090 | 0.80% |
| Provider 27 | 0.98 | 1.12 | 1.12 | $85,490.55 | -0.0065 | -0.58% |
| Provider 41 | 1.07 | 1.13 | 1.12 | $86,685.54 | -0.0070 | -0.62% |
| Provider 36 | 1.04 | 1.14 | 1.14 | $82,723.92 | -0.0052 | -0.45% |
| Provider 33 | 1.02 | 1.15 | 1.14 | $89,318.28 | -0.0083 | -0.72% |
| Provider 37 | 1.04 | 1.18 | 1.17 | $85,086.95 | -0.0063 | -0.53% |
| Provider 66 | 2.03 | 1.39 | 1.38 | $75,167.49 | -0.0016 | -0.12% |

## Provider Group Detailed Results - Commercially Licensed Data

| Provider Group | TCI | RUI (endorsed measure) | RUI (endorsed measure + income) | Average Income | RUI Change | Pct RUI Change |
|---|---|---|---|---|---|---|
| Provider 01 | 0.93 | 0.83 | 0.84 | $ 72,443.51 | 0.0029 | 0.34% |
| Provider 02 | 1.08 | 0.88 | 0.88 | $ 81,736.10 | 0.0017 | 0.19% |
| Provider 03 | 1.14 | 0.88 | 0.89 | $ 77,260.46 | 0.0022 | 0.25% |
| Provider 04 | 1.03 | 0.89 | 0.89 | $ 80,589.91 | 0.0018 | 0.20% |
| Provider 05 | 1.00 | 0.90 | 0.90 | $ 79,052.16 | 0.0020 | 0.22% |
| Provider 06 | 0.85 | 0.90 | 0.90 | $ 90,000.35 | 0.0006 | 0.07% |
| Provider 07 | 0.86 | 0.92 | 0.92 | $ 81,080.58 | 0.0018 | 0.19% |
| Provider 08 | 1.01 | 0.92 | 0.92 | $ 79,498.09 | 0.0020 | 0.21% |
| Provider 09 | 1.41 | 0.92 | 0.93 | $ 81,478.38 | 0.0017 | 0.18% |
| Provider 10 | 1.08 | 0.93 | 0.93 | $ 75,610.49 | 0.0025 | 0.27% |
| Provider 11 | 0.97 | 0.93 | 0.94 | $ 79,045.36 | 0.0020 | 0.22% |
| Provider 12 | 1.62 | 0.94 | 0.94 | $ 75,077.69 | 0.0025 | 0.27% |
| Provider 13 | 1.21 | 0.94 | 0.94 | $ 83,735.62 | 0.0014 | 0.15% |
| Provider 14 | 0.93 | 0.95 | 0.94 | $ 95,896.16 | -0.0002 | -0.02% |
| Provider 15 | 1.19 | 0.95 | 0.95 | $ 82,285.13 | 0.0016 | 0.17% |
| Provider 16 | 0.92 | 0.96 | 0.96 | $ 75,238.24 | 0.0025 | 0.26% |
| Provider 17 | 1.10 | 0.96 | 0.96 | $ 79,490.54 | 0.0020 | 0.20% |
| Provider 18 | 0.95 | 0.96 | 0.96 | $102,194.19 | -0.0010 | -0.10% |
| Provider 19 | 0.85 | 0.97 | 0.97 | $ 74,225.15 | 0.0026 | 0.27% |
| Provider 20 | 1.54 | 0.97 | 0.97 | $ 80,176.59 | 0.0019 | 0.19% |
| Provider 21 | 1.10 | 0.98 | 0.98 | $ 76,567.93 | 0.0023 | 0.24% |
| Provider 22 | 0.93 | 0.98 | 0.98 | $105,426.41 | -0.0014 | -0.14% |
| Provider 23 | 1.32 | 0.98 | 0.99 | $ 72,760.80 | 0.0028 | 0.29% |
| Provider 24 | 0.96 | 0.98 | 0.98 | $114,650.89 | -0.0026 | -0.26% |
| Provider 25 | 0.89 | 0.98 | 0.99 | $ 87,932.23 | 0.0009 | 0.09% |
| Provider 26 | 0.95 | 0.99 | 0.98 | $107,107.13 | -0.0016 | -0.16% |
| Provider 27 | 0.92 | 0.99 | 0.99 | $ 83,708.65 | 0.0014 | 0.14% |

## Provider Group Detailed Results - Commercially Licensed Data - *continued*



| Provider Group | TCI | RUI<br>(endorsed measure) | RUI<br>(endorsed measure + income) | Average Income | RUI Change | Pct RUI Change |
|---|---|---|---|---|---|---|
| Provider 28 | 0.95 | 0.99 | 0.99 | $100,567.76 | -0.0008 | -0.08% |
| Provider 29 | 0.95 | 0.99 | 0.99 | $109,144.48 | -0.0019 | -0.19% |
| Provider 30 | 0.98 | 0.99 | 0.99 | $105,829.78 | -0.0014 | -0.14% |
| Provider 31 | 1.00 | 0.99 | 0.99 | $ 98,557.24 | -0.0005 | -0.05% |
| Provider 32 | 0.92 | 0.99 | 0.99 | $106,951.43 | -0.0016 | -0.16% |
| Provider 33 | 1.00 | 1.00 | 0.99 | $108,508.94 | -0.0018 | -0.18% |
| Provider 34 | 1.13 | 1.00 | 1.00 | $ 77,921.25 | 0.0022 | 0.22% |
| Provider 35 | 0.91 | 1.00 | 1.00 | $ 99,877.19 | -0.0007 | -0.07% |
| Provider 36 | 1.00 | 1.01 | 1.01 | $ 74,293.07 | 0.0026 | 0.26% |
| Provider 37 | 1.31 | 1.01 | 1.02 | $ 82,705.61 | 0.0015 | 0.15% |
| Provider 38 | 1.58 | 1.01 | 1.02 | $ 88,328.20 | 0.0008 | 0.08% |
| Provider 39 | 1.04 | 1.02 | 1.02 | $100,477.23 | -0.0007 | -0.07% |
| Provider 40 | 1.00 | 1.02 | 1.02 | $ 83,196.31 | 0.0015 | 0.15% |
| Provider 41 | 1.19 | 1.03 | 1.03 | $ 91,445.34 | 0.0004 | 0.04% |
| Provider 42 | 1.00 | 1.04 | 1.04 | $103,314.88 | -0.0011 | -0.11% |
| Provider 43 | 1.36 | 1.04 | 1.04 | $ 74,269.13 | 0.0026 | 0.25% |
| Provider 44 | 1.17 | 1.04 | 1.04 | $ 80,904.86 | 0.0018 | 0.17% |
| Provider 45 | 0.96 | 1.05 | 1.04 | $100,032.86 | -0.0007 | -0.07% |
| Provider 46 | 1.21 | 1.05 | 1.05 | $101,489.30 | -0.0009 | -0.08% |
| Provider 47 | 1.08 | 1.06 | 1.06 | $104,618.70 | -0.0013 | -0.12% |
| Provider 48 | 1.09 | 1.06 | 1.06 | $ 83,775.49 | 0.0014 | 0.13% |
| Provider 49 | 1.45 | 1.10 | 1.10 | $ 92,688.65 | 0.0003 | 0.02% |
| Provider 50 | 1.02 | 1.13 | 1.13 | $119,501.39 | -0.0032 | -0.28% |
| Provider 51 | 1.13 | 1.14 | 1.13 | $130,422.56 | -0.0046 | -0.41% |
| Provider 52 | 1.06 | 1.14 | 1.14 | $110,400.00 | -0.0020 | -0.18% |
| Provider 53 | 1.14 | 1.15 | 1.15 | $122,711.45 | -0.0036 | -0.31% |
| Provider 54 | 1.05 | 1.16 | 1.15 | $114,180.16 | -0.0025 | -0.22% |
| Provider 55 | 2.17 | 1.39 | 1.39 | $108,363.80 | -0.0018 | -0.13% |

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps—do not just name a method; what statistical analysis was used*)

Correlations and regression analysis utilized in both validity and the socioeconomic testing papers as well as the results in the Society of Actuaries study indicate that the statistical model used to adjust cost variation is effective. Additionally, because the commercial population's use of the healthcare system is so significantly different from the Medicaid and Medicare populations, through the benefits covered, the predominant conditions treated, and the prices of the services rendered, segmentation is required.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.*

***If stratified, skip to 2b4.9***

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

The Total Resource Use measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). The measure is also limited to commercial only. An evaluation between commercial and Medicaid covered members was also conducted in the socioeconomic testing, highlighting the variation in resource use (results included in 2b4.9.).

The non-risk adjusted Total Cost Relative Resource Values coefficient of 0.88 indicates a high correlation between total resource use and risk score.

| | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **ACG** | **Non-Risk Adj PMPMs** |
| Non-Risk Adj PMPM | 0.62 | 1.00 |
| Non-Risk Adj TCRRVs | 0.88 | 0.78 |
| ACG Risk Adj TCI | 0.03 | 0.79 |
| ACG Risk Adj RUI | 0.14 | 0.45 |
| Price | -0.09 | 0.57 |

| | **R-Sqaured** | |
|---|---|---|
| **Metric** | **ACG** | **Non-Risk Adj PMPMs** |
| Non-Risk Adj PMPM | 0.38 | 1.00 |
| Non-Risk Adj TCRRVs | 0.77 | 0.60 |
| ACG Risk Adj TCI | 0.00 | 0.62 |
| ACG Risk Adj RUI | 0.02 | 0.20 |
| Price | 0.01 | 0.33 |

Validity Paper (see page 5):
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

Socioeconomic Testing Paper (see page 4):
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

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

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

**2b4.9. Results of Risk Stratification Analysis**:

Detailed results can be found on page 4 and 5 of the socioeconomic testing paper:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

Using Census Tract Data the stratification results are shown in the far right column.

| **Model** | **1% Income Increase** | **1% ACG increase** | **Commercial vs. Medicaid Membership** |
|---|---|---|---|
| Total Reimbursement | $ (0.13) | $ 4.22 | $ 133.28 |
| Resource Use | $ 0.16 | $ 4.34 | $ (75.24) |
| Price | $ (0.28) | $ 0.07 | $ 205.36 |

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

Detailed results can be found on page 4 and 5 of the socioeconomic testing paper:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

Product contributed significantly with there being a $133 dollar difference in cost between commercial and Medicaid. The variation in resource use was much less, however, still significant with Medicaid covered members utilizing $75 more dollars of resources. The fact that Medicaid's cost per service is approximately half that of commercial rates drives the differences between the TCOC and Resource Use results.

**2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

---

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Performance is measured on an Index basis relative to 1.00 where each one point (0.01) variation from 1.00 (average) represents a 1% deviation from average. Statistical significance ranges of performance are not necessary as the measure is based on the full population. The results can be analyzed by percentile, percent from mean, standard deviation and clustering methods, this is dependent upon the business application of the measure.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

| | Average ACG Score | | | TCI | | | Price Index | | | Resource Use Index | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Provider Group | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 |
| Provider 01 | 1.11 | 1.09 | 1.09 | 0.87 | 0.84 | 0.84 | 0.89 | 0.91 | 0.89 | 0.98 | 0.93 | 0.93 |
| Provider 02 | ~ | 1.18 | 1.04 | ~ | 0.88 | 0.84 | ~ | 0.90 | 0.88 | ~ | 0.98 | 0.96 |
| Provider 03 | 0.85 | 0.86 | 0.88 | 0.93 | 0.86 | 0.85 | 0.95 | 0.94 | 0.93 | 0.98 | 0.91 | 0.92 |
| Provider 04 | 0.86 | 0.91 | 0.89 | 0.82 | 0.88 | 0.87 | 0.89 | 0.89 | 0.90 | 0.92 | 0.98 | 0.97 |
| Provider 05 | 1.27 | 1.15 | 1.12 | 0.93 | 0.86 | 0.87 | 1.14 | 1.11 | 1.06 | 0.81 | 0.78 | 0.82 |
| Provider 06 | 1.04 | 1.04 | 1.01 | 0.82 | 0.90 | 0.88 | 0.88 | 0.89 | 0.92 | 0.93 | 1.01 | 0.96 |
| Provider 07 | 1.08 | 1.08 | 1.05 | 0.92 | 0.92 | 0.89 | 0.95 | 0.92 | 0.92 | 0.98 | 1.00 | 0.97 |
| Provider 08 | 1.03 | 0.99 | 1.03 | 0.93 | 0.96 | 0.90 | 0.91 | 0.94 | 0.96 | 1.03 | 1.02 | 0.94 |
| Provider 09 | 1.00 | 1.04 | 1.06 | 0.86 | 0.86 | 0.90 | 0.88 | 0.89 | 0.91 | 0.98 | 0.97 | 1.00 |
| Provider 10 | 1.16 | 1.17 | 1.19 | 0.80 | 0.87 | 0.90 | 0.86 | 0.91 | 0.92 | 0.93 | 0.96 | 0.98 |
| Provider 11 | 1.19 | 1.35 | 1.42 | 1.02 | 0.93 | 0.91 | 1.01 | 0.96 | 0.91 | 1.00 | 0.97 | 1.00 |
| Provider 12 | 1.07 | 1.05 | 1.06 | 0.90 | 0.91 | 0.91 | 0.92 | 0.92 | 0.93 | 0.98 | 0.98 | 0.98 |
| Provider 13 | 1.01 | 1.06 | 1.06 | 0.95 | 0.95 | 0.92 | 0.91 | 0.93 | 0.93 | 1.05 | 1.02 | 0.99 |
| Provider 14 | 1.17 | 1.15 | 1.13 | 0.84 | 0.88 | 0.92 | 0.88 | 0.93 | 0.97 | 0.96 | 0.95 | 0.95 |
| Provider 15 | 0.91 | 0.94 | 0.95 | 0.84 | 0.97 | 0.92 | 0.94 | 0.98 | 1.00 | 0.89 | 0.99 | 0.92 |
| Provider 16 | 1.17 | 1.09 | 1.09 | 0.91 | 0.94 | 0.93 | 0.90 | 0.90 | 0.90 | 1.01 | 1.04 | 1.03 |
| Provider 17 | 1.14 | 1.14 | 1.14 | 0.89 | 0.85 | 0.93 | 0.89 | 0.87 | 0.95 | 1.00 | 0.98 | 0.97 |
| Provider 18 | 0.98 | 1.05 | 0.99 | 1.01 | 0.98 | 0.94 | 1.03 | 1.04 | 1.06 | 0.98 | 0.94 | 0.89 |
| Provider 19 | 0.90 | 0.88 | 0.86 | 0.95 | 0.92 | 0.94 | 0.94 | 0.93 | 0.94 | 1.01 | 0.98 | 1.01 |
| Provider 20 | 0.99 | 1.02 | 1.04 | 1.00 | 0.96 | 0.96 | 0.97 | 0.97 | 0.98 | 1.03 | 0.99 | 0.98 |
| Provider 21 | 0.82 | 0.84 | 0.85 | 0.98 | 1.00 | 0.96 | 0.95 | 0.94 | 0.95 | 1.04 | 1.07 | 1.01 |
| Provider 22 | 1.04 | 0.93 | 0.94 | 1.07 | 1.03 | 0.96 | 1.10 | 1.15 | 1.11 | 0.97 | 0.90 | 0.87 |
| Provider 23 | 0.88 | 0.96 | 0.94 | 1.09 | 0.98 | 0.96 | 0.96 | 0.97 | 0.95 | 1.14 | 1.01 | 1.01 |
| Provider 24 | 0.91 | 0.94 | 0.94 | 0.96 | 0.96 | 0.97 | 0.93 | 0.94 | 0.94 | 1.02 | 1.02 | 1.03 |
| Provider 25 | 1.20 | 1.12 | 1.20 | 0.94 | 0.84 | 0.97 | 0.94 | 0.93 | 0.99 | 1.00 | 0.90 | 0.99 |
| Provider 26 | 1.07 | 1.07 | 1.07 | 0.95 | 0.96 | 0.98 | 0.97 | 0.97 | 0.98 | 0.98 | 0.99 | 1.00 |
| Provider 27 | 1.06 | 1.04 | 1.03 | 1.02 | 0.96 | 0.98 | 0.91 | 0.88 | 0.87 | 1.12 | 1.09 | 1.12 |
| Provider 28 | 1.02 | 1.03 | 1.04 | 0.96 | 0.97 | 0.98 | 0.93 | 0.93 | 0.94 | 1.03 | 1.04 | 1.05 |
| Provider 29 | 0.96 | 1.02 | 1.03 | 1.12 | 1.07 | 1.01 | 1.10 | 1.08 | 1.08 | 1.02 | 0.99 | 0.94 |
| Provider 30 | 0.89 | 0.93 | 0.90 | 1.03 | 0.99 | 1.01 | 0.97 | 0.98 | 0.99 | 1.07 | 1.01 | 1.02 |
| Provider 31 | 1.01 | 0.98 | 0.98 | 1.03 | 1.01 | 1.01 | 1.06 | 1.10 | 1.12 | 0.97 | 0.91 | 0.91 |
| Provider 32 | 1.00 | 0.96 | 1.06 | 1.04 | 0.97 | 1.02 | 0.95 | 0.94 | 0.98 | 1.09 | 1.03 | 1.04 |
| Provider 33 | ~ | 0.95 | 1.11 | ~ | 1.00 | 1.02 | ~ | 0.88 | 0.89 | ~ | 1.14 | 1.15 |

The red line divides providers between above and below the average total cost index (1.00).

|  | Average ACG Score |  |  | TCI |  |  | Price Index |  |  | Resource Use Index |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Provider Group | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 |
| Provider 34 | 1.10 | 1.11 | 1.10 | 1.00 | 1.02 | 1.03 | 1.00 | 1.00 | 1.01 | 1.00 | 1.02 | 1.02 |
| Provider 35 | 0.94 | 0.96 | 0.99 | 1.03 | 1.04 | 1.03 | 1.03 | 1.03 | 1.07 | 1.00 | 1.01 | 0.96 |
| Provider 36 | 1.11 | 1.12 | 1.10 | 1.03 | 1.05 | 1.04 | 0.90 | 0.90 | 0.91 | 1.15 | 1.16 | 1.14 |
| Provider 37 | 1.09 | 1.13 | 1.08 | 1.03 | 1.06 | 1.04 | 0.92 | 0.91 | 0.88 | 1.12 | 1.16 | 1.18 |
| Provider 38 | 0.94 | 1.00 | 0.99 | 1.15 | 1.06 | 1.05 | 1.05 | 1.08 | 1.09 | 1.09 | 0.98 | 0.96 |
| Provider 39 | 1.07 | 1.09 | 1.02 | 1.05 | 1.08 | 1.07 | 1.17 | 1.22 | 1.18 | 0.90 | 0.88 | 0.90 |
| Provider 40 | 0.54 | 0.51 | 0.51 | 0.95 | 0.99 | 1.07 | 0.94 | 0.94 | 0.95 | 1.01 | 1.05 | 1.12 |
| Provider 41 | 0.50 | 0.53 | 0.52 | 1.01 | 1.04 | 1.07 | 0.95 | 0.96 | 0.95 | 1.06 | 1.07 | 1.13 |
| Provider 42 | 0.82 | 0.90 | 0.97 | 1.09 | 1.09 | 1.07 | 1.11 | 1.10 | 1.12 | 0.98 | 0.99 | 0.96 |
| Provider 43 | ~ | ~ | 1.07 | ~ | ~ | 1.08 | ~ | ~ | 1.18 | ~ | ~ | 0.91 |
| Provider 44 | 1.12 | 1.06 | 1.09 | 1.13 | 1.09 | 1.10 | 1.12 | 1.09 | 1.08 | 1.01 | 0.99 | 1.02 |
| Provider 45 | 0.88 | 0.88 | 0.90 | 1.25 | 1.20 | 1.11 | 1.25 | 1.28 | 1.25 | 0.99 | 0.93 | 0.88 |
| Provider 46 | 0.92 | 0.90 | 0.87 | 1.10 | 1.15 | 1.11 | 1.16 | 1.21 | 1.22 | 0.95 | 0.95 | 0.91 |
| Provider 47 | ~ | 1.07 | 0.92 | ~ | 1.30 | 1.11 | ~ | 1.18 | 1.23 | ~ | 1.10 | 0.90 |
| Provider 48 | 0.91 | 0.86 | 0.86 | 1.07 | 1.11 | 1.12 | 1.10 | 1.12 | 1.10 | 0.97 | 0.99 | 1.02 |
| Provider 49 | 1.15 | 1.01 | 1.05 | 1.09 | 1.12 | 1.17 | 1.13 | 1.14 | 1.23 | 0.96 | 0.99 | 0.95 |
| Provider 50 | ~ | ~ | 0.97 | ~ | ~ | 1.18 | ~ | ~ | 1.09 | ~ | ~ | 1.08 |
| Provider 51 | 0.83 | 0.79 | 0.84 | 0.95 | 1.00 | 1.19 | 1.10 | 1.10 | 1.13 | 0.86 | 0.91 | 1.05 |
| Provider 52 | 0.98 | 1.09 | 0.99 | 1.36 | 1.31 | 1.19 | 1.36 | 1.32 | 1.34 | 1.00 | 0.99 | 0.89 |
| Provider 53 | 0.85 | 0.92 | 0.90 | 1.20 | 1.26 | 1.20 | 1.23 | 1.23 | 1.25 | 0.98 | 1.03 | 0.96 |
| Provider 54 | 0.89 | 0.97 | 0.96 | 1.36 | 1.23 | 1.20 | 1.28 | 1.31 | 1.30 | 1.06 | 0.94 | 0.92 |
| Provider 55 | 1.13 | 0.92 | 0.90 | 1.19 | 1.38 | 1.25 | 1.32 | 1.36 | 1.27 | 0.90 | 1.02 | 0.98 |
| Provider 56 | 1.02 | 1.03 | 1.04 | 1.31 | 1.29 | 1.26 | 1.25 | 1.23 | 1.23 | 1.05 | 1.05 | 1.02 |
| Provider 57 | ~ | ~ | 0.86 | ~ | ~ | 1.26 | ~ | ~ | 1.34 | ~ | ~ | 0.94 |
| Provider 58 | 0.92 | 1.00 | 0.93 | 1.19 | 1.10 | 1.27 | 1.11 | 1.07 | 1.13 | 1.07 | 1.02 | 1.12 |
| Provider 59 | 0.83 | 0.83 | 0.80 | 1.21 | 1.26 | 1.29 | 1.17 | 1.14 | 1.23 | 1.04 | 1.11 | 1.05 |
| Provider 60 | 0.98 | 0.98 | 1.00 | 1.37 | 1.39 | 1.36 | 1.49 | 1.47 | 1.47 | 0.92 | 0.94 | 0.93 |
| Provider 61 | 0.95 | 0.88 | 0.85 | 1.17 | 1.26 | 1.36 | 1.25 | 1.24 | 1.28 | 0.93 | 1.02 | 1.06 |
| Provider 62 | 0.87 | 0.86 | 0.86 | 1.37 | 1.32 | 1.42 | 1.49 | 1.53 | 1.50 | 0.92 | 0.86 | 0.95 |
| Provider 63 | 0.87 | 0.84 | 0.96 | 1.42 | 1.45 | 1.47 | 1.53 | 1.49 | 1.52 | 0.93 | 0.98 | 0.97 |
| Provider 64 | 1.04 | 1.00 | 0.97 | 1.39 | 1.60 | 1.54 | 1.61 | 1.59 | 1.63 | 0.87 | 1.01 | 0.95 |
| Provider 65 | 1.01 | 1.01 | 0.97 | 1.48 | 1.60 | 1.67 | 1.61 | 1.65 | 1.59 | 0.92 | 0.97 | 1.05 |
| Provider 66 | 1.60 | 1.58 | 1.56 | 1.80 | 1.96 | 2.03 | 1.45 | 1.48 | 1.46 | 1.24 | 1.32 | 1.39 |
| **Network Total** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The Total Resource Use measure can effectively identify variation in performance levels.

Practically meaningful difference in performance will vary by use of the measures. This is because some uses may have a higher threshold for differences. For example, a 10% difference in performance when the result is used for public reporting could be very meaningful in terms of provider patient growth and retention strategies. The same 10% difference may not be as meaningful when using the measures internally for improvement work and identification of a work plan.

The following will give a general sense of the dispersion of the scoring:

Out of the 66 provider groups measured in Total Resource Use:

- 38 were better than average
- 7 were 10% better than average
- 8 were 10% higher than average
- 51 were within 10% of the average

---

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped*.**

**Note**: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps***—***do not just name a method; what statistical analysis was used*)

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

**2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

---

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps***—***do not just name a method; what statistical analysis was used*)

This is a full population-based measure, all data is included in the measure.

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

This is a full population-based measure, all data is included in the measure.

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

This is a full population-based measure, all data is included in the measure.

**See Appendix A for 2012 Testing Submission**

## Feasibility

F.1. **Byproduct of Care Processes**

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

F.1.1. **Data Elements Generated as Byproduct of Care Processes.**
Other
If other: Health Plan Claims data system

F.2. **Electronic Sources**

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

F.2.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*)
ALL data elements are in defined fields in a combination of electronic sources

F.2.1a. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.

F.2.2. **If this is an eMeasure,** provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.

**Attachment:**

F.3. **Data Collection Strategy**

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

F.3.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
Since endorsement we have received some general feedback regarding implementation of the measure. This has helped shape some of the materials and additional testing we've conducted since the measures were first released. HealthPartners has organized a public-facing website with several resources and technical documentation, including toolkits for external organizations to download necessary tools to run the measure, free of charge. In addition, HealthPartners uses SAS to run the measure and not every organization has or uses this software. To address this, HealthPartners organized non-SAS user instructions. By creating these resources and software and putting them in the public domain it has resulted in expanded use. A few users have successfully implemented the NQF-endorsed Resource Use measure according to the specifications, however they are using their previously purchased risk grouper (not ACG).

F.3.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, and algorithm)?**
The measure and software are available free of charge at www.healthpartners.com/tcoc;
The Total Resource Use measure download options are available at:
https://www.healthpartners.com/hp/about/tcoc/toolkit/index.html
The ACG System is widely available within the public and private sectors in the US and abroad.(Bibliography: http://acg.jhsph.org/index.php/resource-center-83/acg-bibliography) The pricing for the ACG System varies for commercial and government entities but is generally based on a per member per year license that is tiered to provide lower per member costs for larger entities. For a commercial plan there is a base fee of $27,000 annually with incremental costs between $0.04 and $0.40 per member per year based on volume, which is inclusive of both license fees and support. Discounted fees are available for government entities and other grant funded not-for-profit entities. Additionally, Johns Hopkins offers research licenses for a very modest cost for academic users incorporating ACGs into published research:
http://www.acg.jhsph.org/index.php?option=com_content&view=article&id=137&Itemid=94

The ACG System technical guide is available here:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf

F.3.3. **If there are any fees associated with the use of this measure as specified, attach the fee schedule here. (Save file as: F3_3_FeeSchedule)**
FeeScheduleTemplate_Proprietary_Fees_V2.0SubmissionForm-Johns_Hopkins_ACG_System_2016-11-636161993265000000.xlsx

## Usability and Use

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

*NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.*

U.1.1. **Current and Planned Use**

| Specific Plan for Use | Current Use (for current use provide URL) |
|---|---|
| | Public Reporting<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf<br><br>Public Health/Disease Surveillance<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/ |

| | |
|---|---|
| | documents/entry_188106.rtf<br><br>Payment Program<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf<br><br>Quality Improvement (external benchmarking to organizations)<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf<br><br>Quality Improvement (Internal to the specific organization)<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf |

U.1.2. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Since endorsement in 2012, uptake of the Total Resource Use measure has expanded across 37 states in the country and used by both national and regional organizations (Coverage). The measure has been used in conjunction with the Total Cost of Care in accountability applications and publicly reported in multiple states for driving improvement.

The following link highlights organizations across the country that have adopted the measure and are currently using the measure for at least one of the uses noted above, including some crossover of multiple uses for some organizations. The URLs of the specific programs are included within the link below to appropriately capture each organization's purpose described in their own words.

Because some of the organizations are using the measure for multiple uses, we have included them based on their predominant category.

https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf

Public Reporting

HealthPartners – Public Reporting, Payment Program, Quality Improvement with Benchmarking

- As a health plan, HealthPartners uses the Total Resource Use measure to incentivize providers to meet Triple Aim goals, optimizing health and patient experience while improving affordability. HealthPartners publicly reports provider group cost results for purposes of transparency for employers, providers, and consumers. The resource results are paired with Total Cost of Care, quality and experience metrics to promote quality improvement with benchmarking across providers.

- HealthPartners has shared savings payment agreements with over 85% of its primary care providers which has increased provider engagement and sharing of appropriate risk as a partnership to lower cost for providers and patients while maintaining quality and experience. Additionally, in conjunction with the Total Cost of Care measure, HealthPartners has begun building upon it by implementing new payment reform models that align incentives among specialists and hospitals to support shared savings with primary care. The new methods include bundled payments for episodes of care as well as models that move away from fee for service and promote coordination of care.

MN Community Measurement – Public Reporting, Quality Improvement with Benchmarking

- In November 2016, MNCM publicly reported Total Resource Use data by provider group in Minnesota using HealthPartners endorsed Total Resource Use measure. Through their multi-stakeholder collaborative process they were able to collect cost data from four health plans and publicly spread the use of the measure to all provider groups in Minnesota, promoting transparency.

Network for Regional Healthcare Improvement – Public Reporting, Quality Improvement with Benchmarking, Quality Improvement

- Eleven regions are part of a project to develop a standardized method of reporting total cost and resource use by using the HealthPartners endorsed measures. During 2015, seven regions were able to share healthcare cost information on over 5 million patients attributed to 20,000 individual physicians through practice level reports. Their work is described in detail in the provided link.

Payment Program

The Alliance

- Utilizes the measures for provider contracting and incentives.

Public Health/Disease Surveillance

The University of Iowa

- Research evaluation for assessing state health system transformation under the State Innovation Model initiative.

Quality Improvement with Benchmarking

Maine Health Management Coalition – regional collaborative

- Commercial premium costs will be measured against benchmarks using the TCOC and Resource Use measures with plans for future public reporting.

Oregon Health Care Quality Corporation – regional collaborative

- In 2015, Q Corp released Clinic Comparison Reports featuring cost, utilization and quality measures to over 150 primary care clinics in Oregon.

HealthInsight and Utah Department of Health, Washington Health Alliance – regional collaboratives

- Regional collaboratives participating in the Network for Regional Healthcare Improvement's project to develop a standardized method of reporting total cost and resource use.

Center for Improving Value in Health Care (CIVHC) – regional collaborative

- Recently began providing results to Colorado primary care groups to help them see how their practice patterns compare.

Midwest Health Initiative – regional collaborative

- Shares data with physician groups and practice sites through community reports with future plans for public reporting.

Quality Improvement

Provider Groups in Minnesota

- Having payment agreements with HealthPartners, several provider groups see the value and are actively engaged in utilizing the Total Resource Use measure. They shared with us how they are using the measure within their own practice and their letters of support are included in the link.

American Hospital Association

- Partnering with HealthPartners to develop a pilot of the measure across their constituents for broader use.

Priority Health

- Evaluating practice efficiencies and pricing fluctuations across Accountable Care Networks.

Providence Health Plans

- Provide efficiency profiling and increasing engagement for improvement and better referral decision making.

Onpoint Health Data

- State organization are utilizing the data for program evaluation.

National Quality Measures Clearinghouse (NQMC) from Agency for Healthcare Research and Quality (AHRQ) reported the following usage between 3/1/15 – 2/29/16

- 1,493 page views for the Total Resource Use Measure

U.1.3. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not applicable

U.1.4. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not applicable

U.2.1. **Progress on Improvement. (Not required for initial endorsement unless available.) Performance results on this measure (current and over time) should be provided in IM.2.2 and IM.2.4.**
**Discuss:**

- **Purpose Progress (trends in performance results)**
- **Geographic area and number and percentage of accountable entities and patients included**

https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf

HealthPartners Medical Group, Park Nicollet Health Services, Essentia Health, CentraCare Health and Fairview are provider groups in Minnesota who are highlighted as engaged users of the measure and who have seen improvement in their care practices. The details they've shared and the strategies they've implemented to lower cost are included in the link provided.

Since endorsement in 2012, there has been a large increase in the number of users who have adopted the Total Resource Use measure, in conjunction with the Total Cost of Care measure, resulting in improvement through greater transparency. An increase in transparency brings an awareness to the rising healthcare costs in our communities and has helped users pinpoint areas for improvement and define strategies to reduce those costs.

HealthPartners has also organized a public-facing website with several resources and technical documentation, including toolkits for external organizations to download the necessary tools to run the measure, free of charge. In addition, HealthPartners has created instructions and toolkits for both SAS and non-SAS users. By creating these resources and software and putting them in the public domain it has resulted in expanded use.

The following link includes details of one specific example demonstrating improvement and features the Northwest Metro Alliance which serves more than 300,000 people receiving care at 9 different clinics and one hospital along with its affiliated specialists. The Alliance's medical cost increases were more than 31 percent lower than the Twin Cities metro average for Commercial patients since they adopted the Total Cost of Care and Resource Use measures in 2010.

Link to and post on website:

https://www.allinahealth.org/uploadedFiles/Content/Customer_Service/Billing_and_insurance/Northwest-Metro-Alliance-5-year-results.pdf

U.2.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not applicable

U.3.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
HealthPartners mitigates risk through the following steps:

- Claims data integrity procedures prior to loading data warehouse through HealthPartners Data Integrity Dept.
- Internal Audit Dept. review of processes & procedures for generating measure
- Provider contracts allow ability to request external audit

•HealthPartners Provider Measurement Policy allows for a 45-day comment period before results are used in any business applications (incentive, public display, etc). Any identified errors ore issues are resolved & corrected

To our knowledge we are not aware of negative unintended consequences from other organizations utilizing the measure.

Since endorsement we have received some general feedback regarding implementation of the measure. This has helped shape some of the materials and additional testing we've conducted since the measure was first released. HealthPartners has organized a public-facing website with several resources and technical documentation, including toolkits for external organizations to download necessary tools to run the measure, free of charge. In addition, HealthPartners uses SAS to run the measure and not every organization has or uses this software. To address this, HealthPartners organized non-SAS user instructions. By creating these resources and software and putting them in the public domain it has resulted in expanded use. A couple of external users have shared feedback on possible barriers with funding of the ACG grouper when the organization has already invested in a different grouper.

## Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

H.1. **Relation to Other NQF-endorsed Measures**
If there are related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

H.1.1. **List of related or competing measures (selected from NQF-endorsed measures)**

H.1.2. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
Not applicable

**H.2. Harmonization**

H.2.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
**Are the measure specifications completely harmonized?**

H.2.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

H.3. **Competing Measure(s)**

H.3.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
**Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)**
Not applicable

## Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** HealthPartners
Co.2 **Point of Contact:** Sue, Knudson, Susan.M.Knudson@healthpartners.com, 952-883-6185-

Co.3 **Measure Developer if different from Measure Steward:** HealthPartners

Co.4 **Point of Contact:** Sue, Knudson, Susan.M.Knudson@healthpartners.com, 952-883-6185-

## Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

List the workgroup/panel members' names and organizations.

Describe the members' role in measure development.

**Measure Developer/Steward Updates and Ongoing Maintenance**

Ad.2 **Year the measure was first released:** 2003

Ad.3 **Month and Year of most recent revision:** 06, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annual

Ad.5 **When is the next scheduled review/update for this measure?** 06, 2017

Ad.6 **Copyright statement:** © 2016 HealthPartners. Reprints allowed for noncommercial purposes only if this copyright notice is prominently included and HealthPartners is given clear attribution as the copyright owner.

Ad.7 **Disclaimers:** Total Cost of Care and Total Resource Use are licensed free of charge with supporting implementation tools at the following website:

www.healthpartners.com/tcoc

Ad.8 **Additional Information/Comments:** HealthPartners public Total Cost of Care and Total Resource Use site: www.healthpartners.com/tcoc

For the purposes of the National Quality Forum Measure Maintenance Review for Endorsed HealthPartners Measures: www.healthpartners.com/tcoc-documents

HealthPartners®

# Appendix A

## Cost and Resource Use Project 2016-2017

National Quality Forum 2012 Measure Endorsement

Total Cost of Care (NQF#1604)

Total Resource Use (NQF#1598)

# 2012 Submission Technical Documentation: Reliability and Validity Testing

HealthPartners' Total Cost of Care and Total Resource Use measures use the same measurement criteria except for the costing method and are considered complementary to each other.

Appendix A supports the Measure Testing Attachments for Total Cost of Care and Total Resource Use Measure Maintenance. The methodology used for both submissions is consistent. Results from the prior testing period using earlier dates of services are included on the following pages.

Results from both testing periods indicate the Total Cost of Care and Total Resource Use measures are both reliable and valid.

HealthPartners®

HealthPartners Technical Documentation

# Total Cost of Care
# Bootstrap Reliability Analysis

## Purpose

Determine the reliability of the Total Cost of Care (TCI) measure.

## Table of Contents

## *Overview of Analysis*

Total Cost of Care (TCI) is a measure of a provider's effectiveness of managing their primary care attributed population across the care continuum. The TCI measure was applied to HealthPartners' primary care metro providers as per the measure specifications and results were calculated for 2007, 2008, and 2009.

The reliability testing demonstrates the repeatability of producing the same results a high proportion of the time. To measure the reliability of the TCI measure a 90% random sample and a bootstrapping technique were employed. In these methods, reliability is measured as the mean of the variance between sampling iterations and the actual results.

In addition, the TCI measure was analyzed over time to demonstrate stability and sensitivity to provider changes or improvement initiatives.

These methods were chosen as they represent the measure intent, which is that the TCI measure represents providers' average total cost of care across their population. Since the measure is aggregated to the provider group level there is no need to quantify the variability at the member level into the evaluation.

In the 90% random sample method, the members that were attributed to a provider group were randomly sampled at the 90% membership level without replacement. This technique was employed to simulate variation within a provider group by leveraging their own population and case-mix. This method gives an indication as to the repeatability of the measure by comparing how closely the actual total cost measure is to the 90% sampled averages and simulates any potential member selection bias.

In the bootstrapping method members that were attributed to a provider group were randomly selected with replacement. This method maximizes variation around a provider group's total cost of care as each randomly selected iteration (sample populations) does not truly represent the provider's case mix of patients. This method was performed in the same fashion as above to support and validate the results found in the 90% sample method.

## *Overall Conclusions*

- The differences between provider Actual TCI results and both the 90% sample and bootstrap mean results are very small.
  - Ranging from -0.0069 to 0.00083 in the 90% sample in 2009.
  - Ranging from -0.00067 to 0.00252 in the bootstrap in 2009.
  - These results indicate that the TCIs for each provider group are repeatable and consistent.
- A provider's performance is relatively consistent across all three years with an average difference of 0.031.
  - These differences in provider performance over time occur because of known changes in fee schedules, collaborating provider usage and resource use saving initiatives can account for the differences.
  - Since the measure is designed to capture and reflect changes in these areas, we expect to see some explainable variability within a provider group over time.

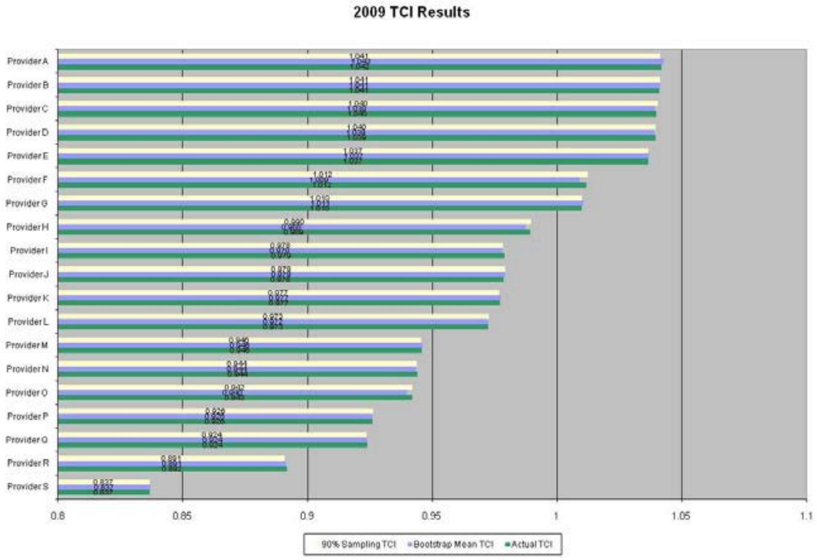
## *Methodology*

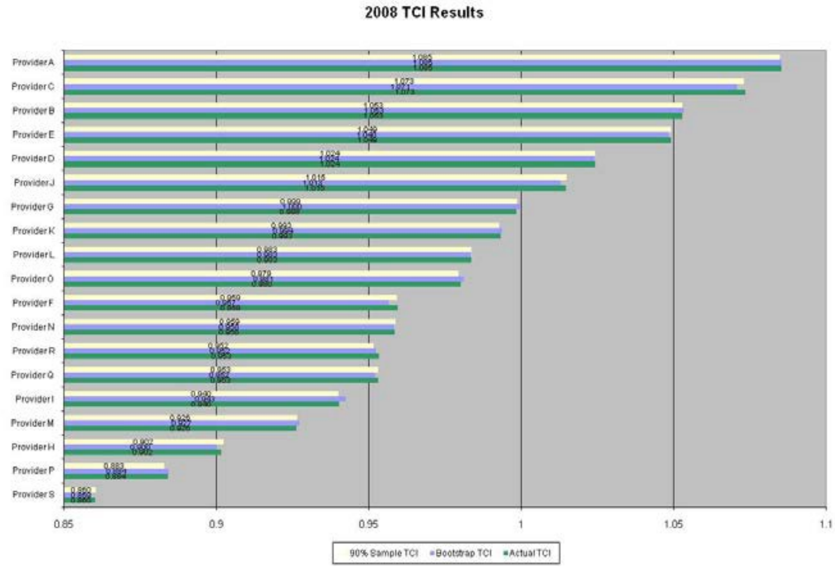
In the 90% sample method, 90% of attributed provider group members were randomly selected, without replacement. A 90% sample was used despite having the full health plan provider population, to simulate any potential member selection bias. The sampling process was performed using the SAS PROC SURVEYSELECT procedure with the Simple Random Sample (SRS) option. This method allows for each attributed member to be selected only one time until 90% of the total provider population has been reached. The 90% sampling process was repeated 500 times for each provider group and year analyzed. Attributed members' total costs were aggregated in each sample to produce 500 TCI results for each provider group for each year (see Figure 1 in the definitions section for more information). Once the 500 samples were created for each provider group, the total costs of care of each sample for each provider group were compared to the metro average to produce risk adjusted indices. The Total Cost indices from each of the sampling iterations for each provider group/year were then compared to the actual TCI indices for each provider group/year and the mean variance was computed.

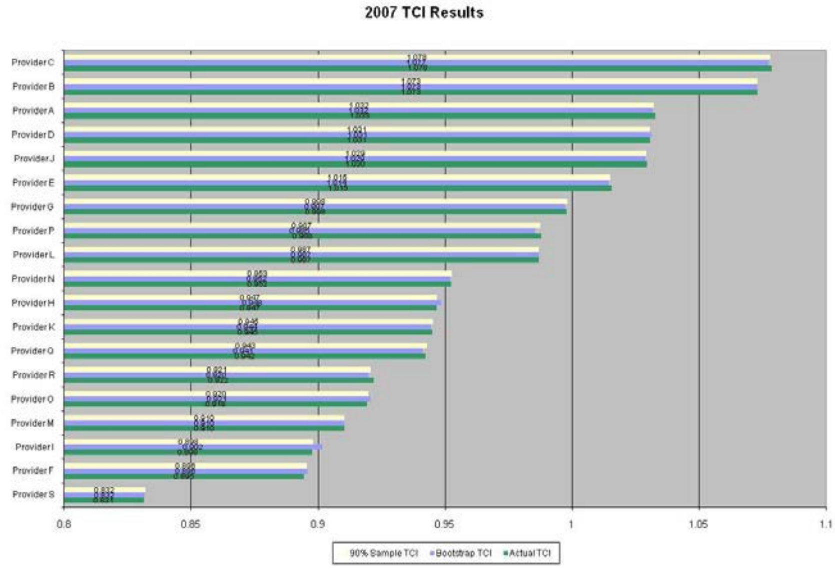
To perform the bootstrap, the SAS PROC SURVEYSELECT procedure with the Unrestricted Random Sample option for full replacement utilized to create a series of random samples for each provider group being measured. Full replacement means that one observation is drawn at random, recorded, and then placed back into the data pool so that it can be drawn again if randomly selected. The numbers of records sampled are drawn such that the samples created are the same size as the original number of attributed members for the provider group. In this way, it is theoretically possible (although virtually improbable) to produce a sample of size n that could consist of the same record drawn n times in a row. This was done to artificially maximize the variance within the defined populations. This sample process was performed 500 times for each year and provider group being analyzed, to produce 500 sets of risk adjusted Total Cost of Care results for each provider for each year (see Figure 2 in the definitions section for more information). The Total Cost indices from each of the sampling iterations for each provider group/year were then compared to the actual TCI indices for each provider group/year and the mean variance was computed.

## Bootstrap and 90% Random Sample

The mean TCI results from the bootstrap and 90% samples compared to the actual TCI results for each provider group and year are displayed in the tables and graphs on the following pages.







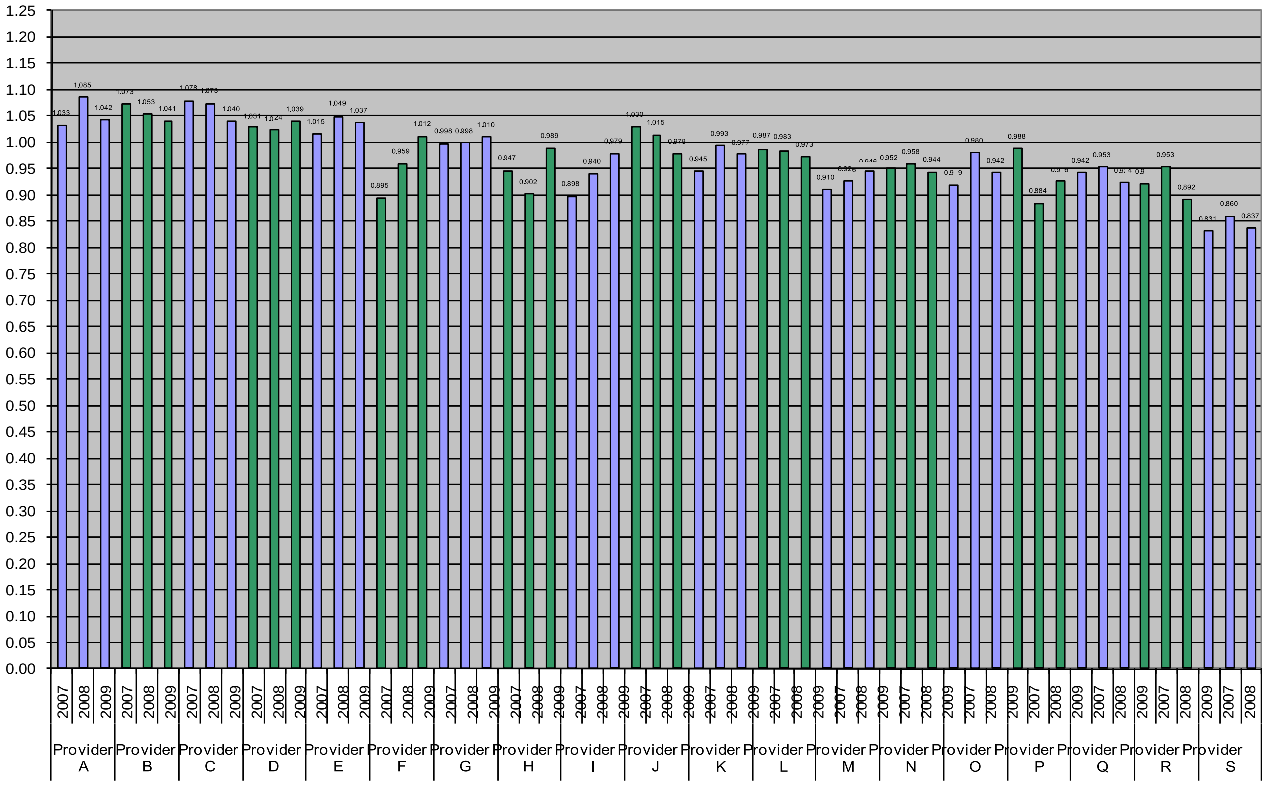
## *Bootstrap and 90% Random Sample Results*

- The differences between provider Actual TCI results and both the 90% sample and bootstrap mean results are very small ranging from -0.0069 to 0.00083 in the 90% sample to -0.00067 to 0.00252 in the bootstrap in 2009.
- The results indicate that the TCIs for each provider group are repeatable and consistent.

## TCI Consistency Over Time

The TCI results are displayed from 2007 through 2009 for the HealthPartners Primary Care Metro Network. The measure differentiates between providers however they remain relatively consistent over time. The factors that drive variation between years within a provider are cost per unit control and resource use management.

Provider Actual TCI Over Time



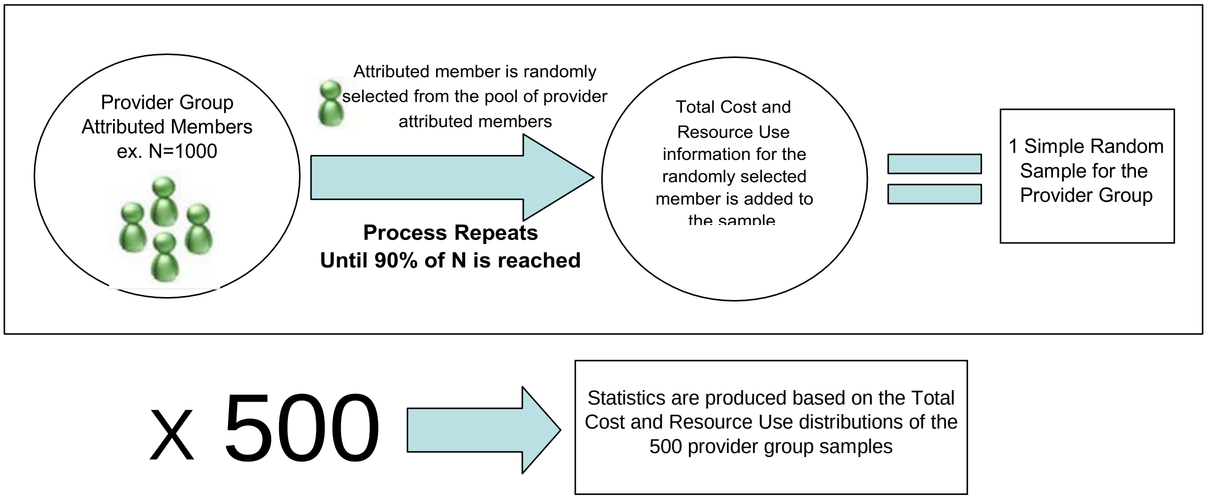
## TCI Consistency Over Time Results

A provider's relative performance is relatively consistent across all three years with an average difference of 0.031.

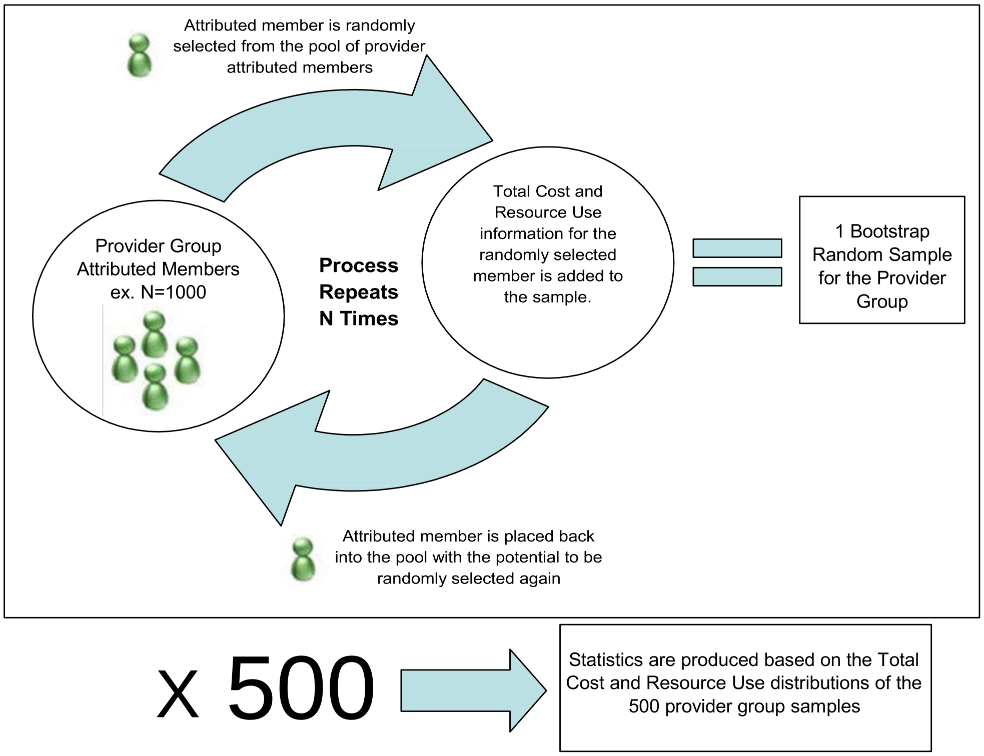
- These differences in provider performance over time occur because of known changes in fee schedules, collaborating provider usage and resource use saving initiatives can account for the differences.
- Since the measure is designed to capture and reflect changes in these areas, we expect to see some explainable variability within a provider group over time.

## *Definitions and Examples*

### Figure 1: 90% Sampling – Simple Random Sample Without Replacement


Provider Group
Attributed Members
ex. N=1000
Attributed member is randomly selected from the pool of provider attributed members
Process Repeats
Until 90% of N is reached
Total Cost and Resource Use information for the randomly selected member is added to the sample
1 Simple Random Sample for the Provider Group
x 500
Statistics are produced based on the Total Cost and Resource Use distributions of the 500 provider group samples

### Figure 2: Bootstrap Sampling – Unrestricted Random Sampling With Full Replacement


Attributed member is randomly selected from the pool of provider attributed members
Provider Group
Attributed Members
ex. N=1000
Process
Repeats
N Times
Total Cost and Resource Use information for the randomly selected member is added to the sample.
1 Bootstrap Random Sample for the Provider Group
Attributed member is placed back into the pool with the potential to be randomly selected again
x 500
Statistics are produced based on the Total Cost and Resource Use distributions of the 500 provider group samples

HealthPartners®

HealthPartners Technical Documentation

# Total Resource Use
# Bootstrap Reliability Analysis

## Purpose

Determine the reliability of the Resource Use Index (RUI) measure.

## Table of Contents

## Overview of Analysis

Resource Use Index (RUI) is a measure of a provider's effectiveness of managing their primary care attributed population across the care continuum. The RUI measure was applied to HealthPartners primary care metro providers as per the measure specifications and results were calculated for 2007, 2008, and 2009.

The reliability testing demonstrates the repeatability of producing the same results a high proportion of the time. To measure the reliability of the RUI measure a 90% random sample and a bootstrapping technique were employed. In these methods, reliability is measured as the mean of the variance between sampling iterations and the actual results.

In addition, the RUI measure was analyzed over time to demonstrate stability and sensitivity to provider changes or improvement initiatives.

These methods were chosen as they represent the measure intent, which is that the RUI measure represents providers' average resource use across their population. Since the measure is aggregated to the provider group level there is no need to quantify the variability at the member level into the evaluation.

In the 90% random sample method, the members that were attributed to a provider group were randomly sampled at the 90% membership level without replacement. This technique was employed to simulate variation within a provider group by leveraging their own population and case-mix. This method gives an indication as to the repeatability of the measure by comparing how closely the actual resource use measure is to the 90% sampled average and simulates any potential member selection bias.

In the bootstrapping method members that were attributed to a provider group were randomly selected with replacement. This method maximizes variation around a provider group's resource use as each randomly selected iteration (sample populations) does not truly represent the provider's case mix of patients. This method was performed in the same fashion as above to support and validate the results found in the 90% sample method.

## Overall Conclusions

- The differences between provider Actual RUI results and both the 90% sample and bootstrap mean results are very small.
    - Ranging from -0.00449 to 0.00125 in the 90% sample in 2009.
    - Ranging from -0.00473 to 0.00105 in the bootstrap in 2009.
    - These results indicate that the RUIs for each provider group are repeatable and consistent.
- A provider's performance is relatively consistent across all three years with an average difference in RUI between 2008 and 2009 of 0.0125.
    - These differences in provider performance over time occur because of known changes in fee schedules, collaborating provider usage and resource use saving initiatives can account for the differences.
    - Since the measure is designed to capture and reflect changes in these areas, we expect to see some explainable variability within a provider group over time.

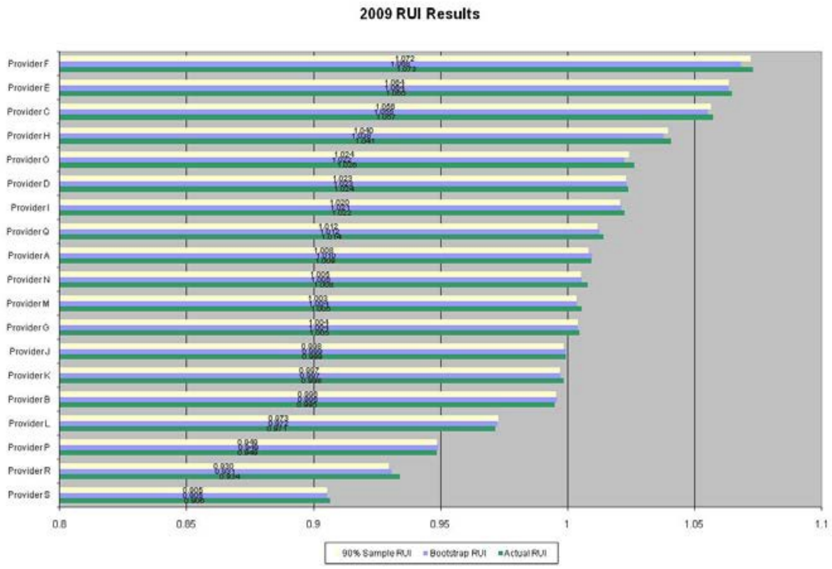
## *Methodology*

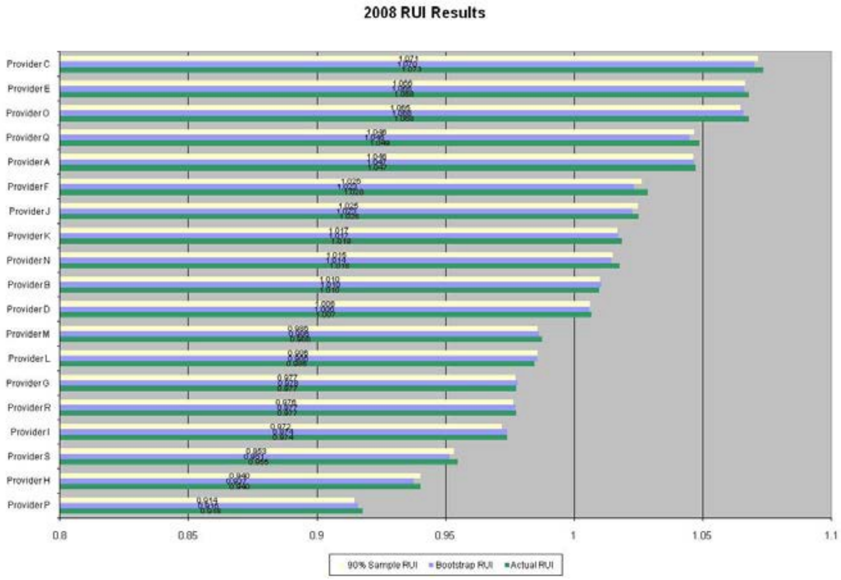
In the 90% sample method, 90% of attributed provider group members were randomly selected, without replacement. A 90% sample was used despite having the full health plan provider population, to simulate any potential member selection bias. The sampling process was performed using the SAS PROC SURVEYSELECT procedure with the Simple Random Sample (SRS) option. This method allows for each attributed member to be selected only one time until 90% of the total provider population has been reached. The 90% sampling process was repeated 500 times for each provider group and year analyzed. Attributed members' resource use was aggregated in each sample to produce 500 RUI results for each provider group for each year (see Figure 1 in the definitions section for more information). Once the 500 samples were created for each provider group, the resource use of each sample for each provider group was compared to the metro average to produce a risk adjusted index. The Resource Use Index from each of the sampling iterations for each provider group/year was then compared to the actual RUI for each provider group/year and the mean variance was computed.

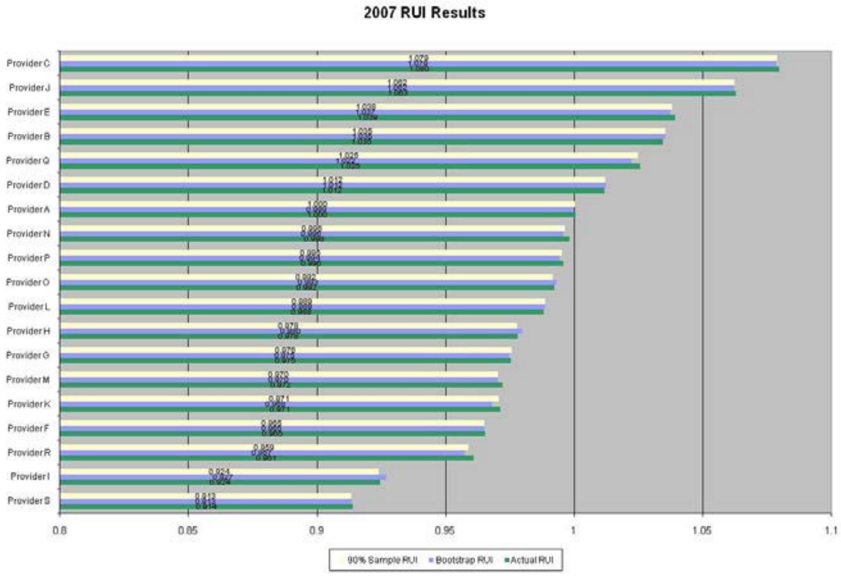
To perform the bootstrap, the SAS PROC SURVEYSELECT procedure with the Unrestricted Random Sample option for full replacement utilized to create a series of random samples for each provider group being measured. Full replacement means that one observation is drawn at random, recorded, and then placed back into the data pool so that it can be drawn again if randomly selected. The numbers of records sampled are drawn such that the samples created are the same size as the original number of attributed members for the provider group. In this way, it is theoretically possible (although virtually improbable) to produce a sample of size n that could consist of the same record drawn n times in a row. This was done to artificially maximize the variance within the defined populations. This sample process was performed 500 times for each year and provider group being analyzed, to produce 500 sets of risk adjusted Resource Use results for each provider for each year (see Figure 2 in the definitions section for more information). The Resource Use Index from each of the sampling iterations for each provider group/year was then compared to the actual RUI for each provider group/year and the mean variance was computed.

## *Bootstrap and 90% Random Sample*

The mean Resource Use result from the bootstrap and 90% samples compared to the actual Resource Use result for each provider group and year is displayed in the tables and graphs on the following pages.







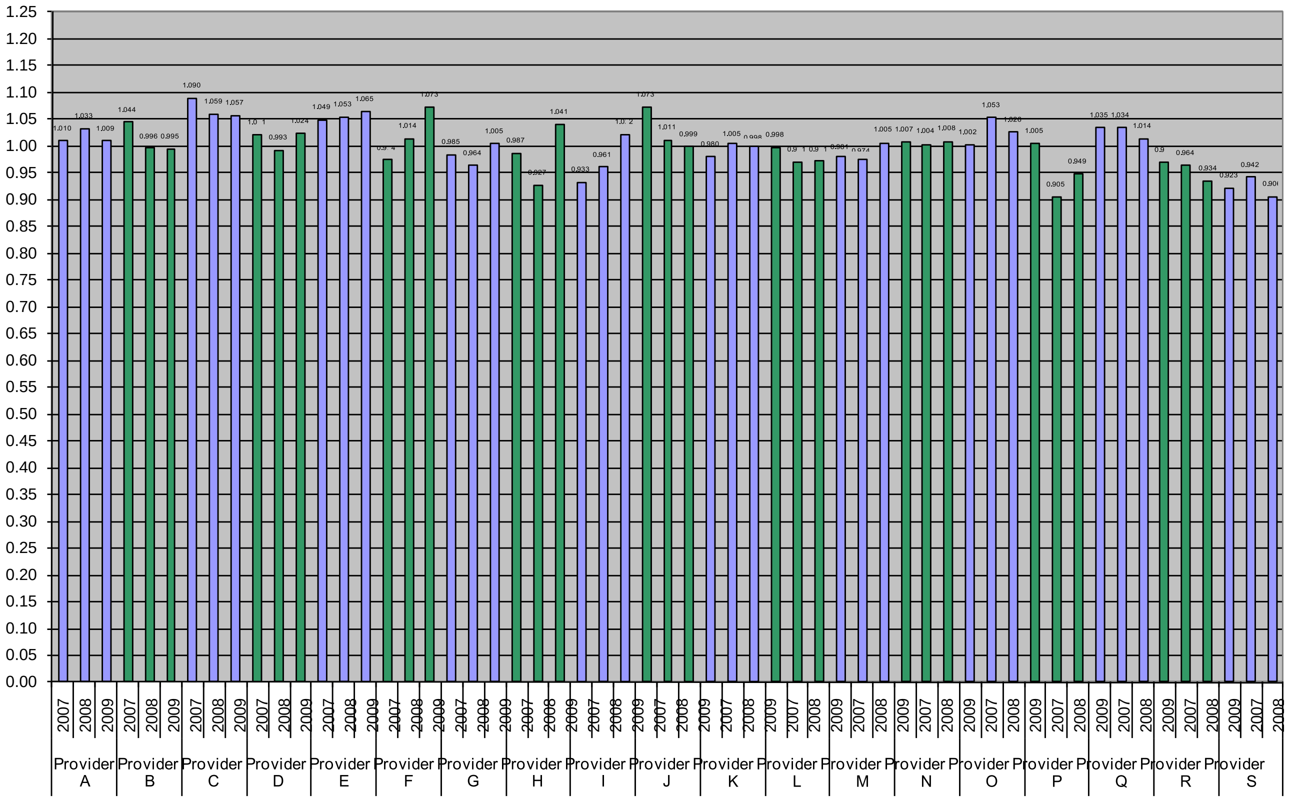
## Bootstrap and 90% Random Sample Results

- The differences between provider Actual RUI results and both the 90% sample and bootstrap mean results are very small ranging from -0.00449 to 0.00125 in the 90% sample to -0.00473 to 0.00105 in the bootstrap in 2009.
- The results indicate that the RUIs for each provider group are repeatable and consistent.

## RUI Consistency Over Time

The Resource Use results are displayed from 2007 through 2009 for the HealthPartners Primary Care Metro Network. The measure differentiates between providers however they remain relatively consistent over time. The factor that drives variation between years within a provider is resource use management.

Provider Actual RUI Over Time



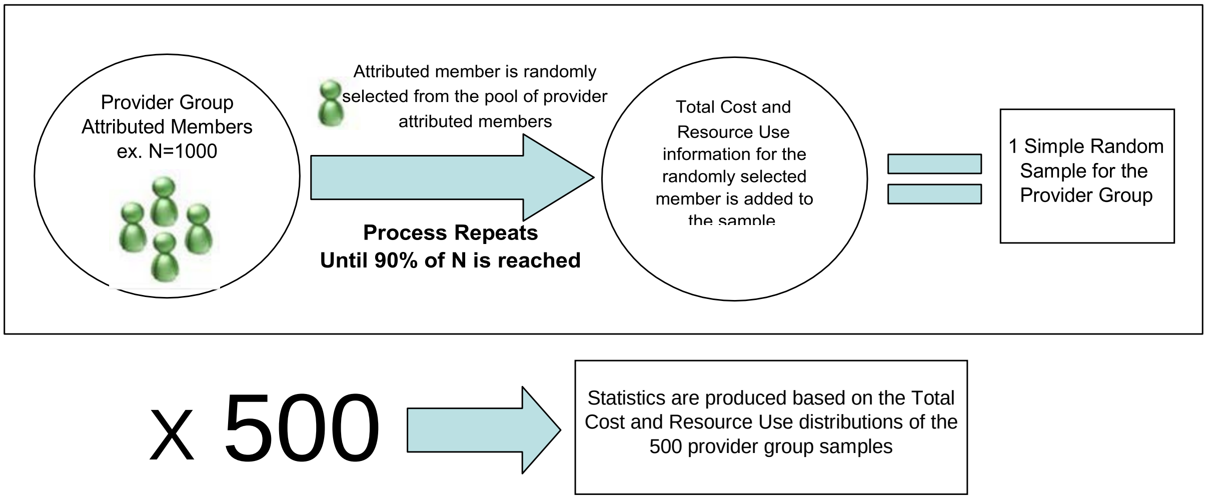
## RUI Consistency Over Time Results

A provider's relative performance is relatively consistent across all three years with an average difference of 0.0125.

- These differences in provider performance over time occur because of known changes in fee schedules, collaborating provider usage and resource use saving initiatives can account for the differences.
- Since the measure is designed to capture and reflect changes in these areas, we expect to see some explainable variability within a provider group over time.

## *Definitions and Examples*

### Figure 1: 90% Sampling – Simple Random Sample Without Replacement



Provider Group Attributed Members ex. N=1000

Attributed member is randomly selected from the pool of provider attributed members

Total Cost and Resource Use information for the randomly selected member is added to the sample

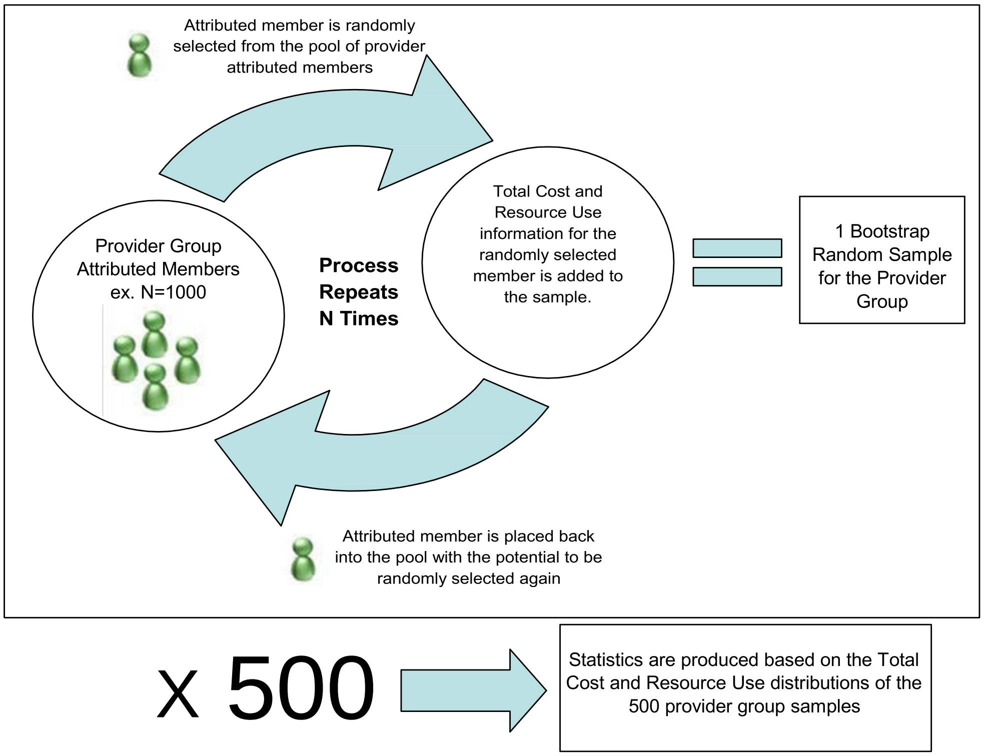
1 Simple Random Sample for the Provider Group

**Process Repeats Until 90% of N is reached**

x 500

Statistics are produced based on the Total Cost and Resource Use distributions of the 500 provider group samples

### Figure 2: Bootstrap Sampling – Unrestricted Random Sampling With Full Replacement



Attributed member is randomly selected from the pool of provider attributed members

Provider Group Attributed Members ex. N=1000

**Process Repeats N Times**

Total Cost and Resource Use information for the randomly selected member is added to the sample.

1 Bootstrap Random Sample for the Provider Group

Attributed member is placed back into the pool with the potential to be randomly selected again

x 500

Statistics are produced based on the Total Cost and Resource Use distributions of the 500 provider group samples

HealthPartners®

HealthPartners Technical Documentation

# Total Cost of Care and Total Resource Use Validity Testing Analysis

## Purpose

To evaluate the Total Cost of Care and Resource Use measures by comparing the findings and correlations to other known information sources and metrics to determine the validity of the measures.

## Table of Contents

## *Overview of Analysis*

The Total Cost of Care and Resource Use are measures of a provider's risk adjusted cost and resource use effectiveness at managing their primary care attributed population across the care continuum. The Total Cost of Care and Resource Use measures were applied to HealthPartners primary care metro providers as per the specifications of the measures. Additional standard utilization metrics were also applied to the underlying data in the actual and risk adjusted forms. The total cost index (TCI) and total resource use index (RUI) findings are compared by provider group to the actual and risk adjusted utilization metrics to determine the correctness of conclusions.

## Methodology

The Total Cost of Care and Resource Use measures should differentiate between providers based on the cost per member and/or consumption of resources per member given all other factors are equal. The ACG adjustment controls for variations in the illness burden of the patients and the peer grouping controls for various patient demographics, provider types and types of product. The remaining factors reflect what the provider can control.

The Total Cost of Care and Resource Use measures should show various strengths of correlations to known utilization metrics. These correlation strengths will depend upon how fully encompassing the utilization metric is within the component being measured and whether the metrics are risk adjusted. For example the admit count utilization measure should be highly correlated to the inpatient resource use as the only factor not accounted for in the admit count measure is intensity (aka: level of treatment). When risk adjustment is applied the correlation will be reduced as the illness burden variation is removed.

The Total Cost of Care and Resource Use measures are designed to evaluate the entire patient and/or provider. Since a person centered measure does not currently exist, the utilization metrics are being used as a proxy to evaluate the correctness and accuracy of the conclusions drawn by the Total Cost of Care and Resource Use measures. These comparisons and correlations should be considered as directional and are not absolute. The utilization metrics do not measure intensity or cost per unit and are targeted to measure a specific service therefore the correlations to the Total Cost of Care and Resource Use need interpretation as high correlation are not always the ultimate goal or the expected result.

## Analysis Overview

- The Pearson correlation coefficients are calculated at the network level between provider groups. In general, the correlation coefficient is an indicator of the level of connection or influence two measures have on each another.
- The correlation coefficient scores range from negative one to positive, with the closer to either value indicating the more influence or connection and the close to zero indicating no influence.
- When the correlation is positive both values move in the same direction and when the correlation is negative the values move in the opposite direction.
    - Positive correlation example: the more admits that are incurred, the more total spend is accumulated. In this case the correlation coefficient would be close to 1.0.
- Network Overview Non Risk Adjusted Metrics
    - Correlations between the ACG score and the non-ACG adjusted cost PMPM and TCRRV PMPM.
    - Correlations between known utilization metrics and the ACG score and the non-ACG adjusted cost PMPM and TCRRV PMPM.
    - Correlations between known utilization metrics within specific places of service and the non-ACG adjusted cost PMPM and TCRRV PMPM for the corresponding places of service. .

- Network Overview Risk Adjusted Metrics
  - Correlations between the ACG score and the Total Cost Index and Resource Use Index.
  - Correlations between known utilization metrics and the overall TCI and RUI.
  - Correlations between known utilization metrics within specific places of service and the TCI and RUI for the corresponding places of service Rx only has a TCI as there is no price variation between providers for pharmacy services.

## Member Population

- Members age 1 – 64 included (babies < 1 and members age 65+ are excluded).
- Members are included if they are enrolled for a minimum of 9 months during the 12 month claims window.
- Commercial products only.
- Attributed members only.
- A member is assigned to the provider group that provides the largest percentage of the primary care office visits.
- In the event of a tie, the provider group with the most recent visit is attributed the member.
- Members that do not have a primary care office visit are excluded from attribution and TCOC.
- Metro Primary Care Providers with more than 600 members that meet the above criteria.

## Network Analysis Overview

- HealthPartners primary care metro network consists of 19 individual provider groups that have 230 clinic sites.
- The total membership of the primary care attributed metro network is over 300,000 members in 2009.
- The variations between provider groups within the following metrics:
  - ACG score variation – 0.85 points (min 0.73 and max 1.59).
  - Total Cost of Care variation – 0.21 points (min 0.84 and max 1.04).
  - Resource Use variation – 0.16 points (min 0.91 and max 1.07).
  - Provider group size vary from 600 to 100,000 members.

## Metrics

- Total Cost Index – TCI: a provider's ACG Adjusted total cost per member per month divided by the metro average ACG Adjusted total cost per member per month.
- Total Care Relative Resource Use Value Index – RUI: a provider's ACG Adjusted total resource use per member per month divided by the metro average ACG Adjusted total resource use per member per month.
  - The Total Care Relative Resource Use Values (TCRRVs) place a relative value unit on all health care services and are the basis of the resource use index (see TCRRV documentation on www.healthpartners.com/tcoc).
- Price Index – PI: a natural byproduct of the TCI and RUI. By definition the only variance between the TCI and RUI is that RUI is void of price.
- Each of these measures is repeated for the four major places of service, inpatient, outpatient, professional and pharmacy.

- Utilization metric indices are counts of distinct services compared to the peer group average.
  - These utilization metrics are risk adjusted through the ACG methodology, which is accomplished by creating expected value by ACG cell.

## *Overview of Conclusions*

- The Total Cost of Care and the Resource Use measures accurately and consistently identified providers that are low or high performers as the measures were able to evaluate a provider's cost and resource effectiveness as supported by known utilization measures.
- There is a high correlation between ACG score and the unadjusted PMPM and TCRRVs which indicates that the Actual PMPM and the Actual TCRRVs are a good measure of the consumption of resources.
- The ACGs, Actual PMPMs and Actual TCRRVs have similar correlation scores to all utilization metrics which indicate the TCRRVs are performing as expected and are a solid measure of resource consumption.
- The Resource Use measure has a high correlation (0.77) to a composite utilization index, which was developed as a proxy to measure total resource consumption (see RUI vs. Risk adjusted Composite Utilization Index section).
- The Total Cost of Care and Resource Use measures differentiate between provider groups accurately and correctly as supported by a wide array of utilization metrics (see Detailed Provider to Provider Analysis and Detailed Provider to Provider – Selected Place of Service sections).

## *Total Cost of Care & Resource Use Report*

This graphic is displayed for a frame of reference. Each provider group has an ACG index, Total Cost Index and a Resource Use Index and each of these are relative to the metro total. The red line divides providers between above and below the average total cost index. There are also utilization metrics described in the Metric Overview section that are calculated for each provider group that are shown later in the analysis.

## Primary Care Provider Network Overview

Commercial, Continuously Enrolled, Excluding Babies and 65+
Dates of Service within each Year
Indexed to the Metro Average

| | Average ACG Score | | | TCI | | | Resource Use Index | | |
|---|---|---|---|---|---|---|---|---|---|
| **Provider Group** | **2007** | **2008** | **2009** | **2007** | **2008** | **2009** | **2007** | **2008** | **2009** |
| Provider O | 0.98 | 0.96 | 0.96 | 0.83 | 0.86 | 0.84 | 0.91 | 0.95 | 0.91 |
| Provider G | 1.03 | 1.16 | 1.09 | 0.92 | 0.95 | 0.89 | 0.96 | 0.98 | 0.93 |
| Provider M | 1.07 | 1.04 | 1.09 | 0.94 | 0.95 | 0.92 | 1.03 | 1.05 | 1.01 |
| Provider D | 1.02 | 1.03 | 1.03 | 0.99 | 0.88 | 0.93 | 1.00 | 0.92 | 0.95 |
| Provider N | 1.04 | 1.05 | 1.04 | 0.92 | 0.98 | 0.94 | 0.99 | 1.07 | 1.03 |
| Provider F | 1.06 | 1.06 | 1.05 | 0.95 | 0.96 | 0.94 | 1.00 | 1.02 | 1.01 |
| Provider S | 0.94 | 0.92 | 0.92 | 0.91 | 0.93 | 0.95 | 0.97 | 0.99 | 1.01 |
| Provider I | 1.01 | 1.02 | 1.02 | 0.99 | 0.98 | 0.97 | 0.99 | 0.98 | 0.97 |
| Provider Q | 0.90 | 0.92 | 0.97 | 0.94 | 0.99 | 0.98 | 0.97 | 1.02 | 1.00 |
| Provider K | 0.77 | 0.79 | 0.79 | 1.03 | 1.01 | 0.98 | 1.06 | 1.03 | 1.00 |
| Provider L | 0.95 | 0.94 | 0.95 | 0.90 | 0.94 | 0.98 | 0.92 | 0.97 | 1.02 |
| Provider B | 0.93 | 0.94 | 1.00 | 0.95 | 0.90 | 0.99 | 0.98 | 0.94 | 1.04 |
| Provider E | 1.03 | 1.00 | 0.99 | 1.00 | 1.00 | 1.01 | 0.98 | 0.98 | 1.00 |
| Provider R | 1.07 | 1.05 | 1.03 | 0.89 | 0.96 | 1.01 | 0.97 | 1.03 | 1.07 |
| Provider H | 1.01 | 0.96 | 1.00 | 1.02 | 1.05 | 1.04 | 1.04 | 1.07 | 1.06 |
| Provider A | 1.01 | 1.03 | 1.02 | 1.03 | 1.02 | 1.04 | 1.01 | 1.01 | 1.02 |
| Provider C | 0.75 | 0.76 | 0.73 | 1.08 | 1.07 | 1.04 | 1.08 | 1.07 | 1.06 |
| Provider P | 0.96 | 0.95 | 0.94 | 1.07 | 1.05 | 1.04 | 1.03 | 1.01 | 0.99 |
| Provider J | 1.64 | 1.61 | 1.59 | 1.03 | 1.09 | 1.04 | 1.00 | 1.05 | 1.01 |
| **Metro Total** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |

## *Correlations Overview*

### Correlations Between ACG Score, Actual PMPMs, Actual TCRRVs, Risk Adj TCI, and Risk Adj RUI

Since the ACG is an industry standard tool that measures resource use there should be strong correlations between it and the Actual PMPMs and Actual TCRRVs. The Actual PMPM and TCRRV correlations should be similar to the ACG score; however the TCRRV's correlation should be stronger as the Actual PMPMs are not a true unbiased measure of resources, as it is impacted by the unit cost of each of the providers within the analysis.

| Non-Risk Adjusted | Correlation Coefficient | |
|---|---|---|
| **Metric** | **ACG** | **Actual PMPMs** |
| Actual PMPM | 0.95 | |
| Actual TCRRVs | 0.97 | 0.98 |
| ACG Risk Adj TCI | 0.06 | 0.37 |
| ACG Adjusted RUI | -0.09 | 0.15 |

- There is a high correlation between ACG score and the unadjusted PMPM and TCRRVs which indicates that the Actual PMPM and the Actual TCRRVs are a good measure of the consumption of resources.
- There is a low correlation between ACG score and the risk adjusted TCI and RUI. This would indicate that a provider can have a high or low ACG score and still have a high or low risk adjusted TCI.
- There is a lower correlation between the risk adjusted RUI and Actual PMPMs than the risk adjusted TCI and Actual PMPMs as the risk adjusted RUIs are not impacted by the cost per unit.

## *Non-Risk Adjusted Correlations*

### Correlations Between ACG Score, Actual PMPMs, and Actual TCRRVs to Non-Risk Adj Utilization Metrics

Since the ACG score, Actual PMPMs and Actual TCRRVs are a measure of the consumption of health care services, there should be some correlation between these values and known utilization metrics. These correlations will not be absolute as the utilization metrics encompass only a portion of the total member's experience. It is expected however that the Actual TCRRVs, which is the underlying value that measures resource use, should have similar correlations to the Actual PMPMs and ACG scores.

- The ACGs, Actual PMPMs and Actual TCRRVs have similar correlation scores which indicate the TCRRVs are performing as expected and are a solid measure of resource consumption.
- There is a high correlation between ACG score, Actual PMPM and Actual TCRRVs to the prescriptions per 1,000 and E&Ms per 1,000. Since E&M visits and Rx scripts are a good indicator of member utilization and total health care consumption it is a positive sign that there is a strong correlation to the ACGs, actual PMPMs and TCRRVs.
- The admits per 1,000 and ER per 1,000 have the lowest correlations to the ACG and actual PMPMs which would indicate that these are low volume service and are outcome based measures.

## Correlation Between the Non-Risk Adjusted Place of Service Metrics and Actual PMPMs &Actual TCRRVs

There should be a correlation between the place of service utilization metrics and the Actual PMPMs and TCRRVs of the corresponding place of service. The magnitude of the correlation is dependent upon the utilization metric's penetration within the place of service and the cost and/or resource intensity of the metric. The Actual PMPMs correlation to the utilization metric will also be impacted by the unit cost of each of the providers within the analysis.

## Inpatient Utilization Correlation to the Inpatient Actual PMPMs & Actual TCRRVs:Non-Risk Adjusted

There should be strong correlations between the admit rate to the Actual PMPMs and Actual TCRRVs as the only two factors not measured by the admits are the intensity and unit cost of the services performed.

| **Inpatient** | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **IP Actual PMPMs** | **IP Actual TCRRVs** |
| Admits/1000 | 0.87 | 0.88 |

## Outpatient Utilization Correlation to the Outpatient Actual PMPMs & Actual TCRRVs:Non-Risk Adjusted

There should be solid correlations between the ER and outpatient surgery rates to the Actual PMPMs and Actual TCRRVs as these two utilization metrics combine to encompass approximately 50% of the total outpatient spend.

| **Outpatient** | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **OP Actual PMPMs** | **OP Actual TCRRVs** |
| ER/1000 | 0.85 | 0.78 |
| OP Surgery/1000 | 0.68 | 0.77 |

## Professional Utilization Correlation to the Professional Actual PMPMs & Actual TCRRVs:Non-Risk Adjusted

There should be solid correlations between the E&M visits and Lab/Path services to the Actual PMPMs and Actual TCRRVs as they represent 45% of the professional spend, but they also are good indicators of patients that consume medical services.

| **Professional** | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **Prof Actual PMPMs** | **Prof Actual TCRRVs** |
| E&M/1000 | 0.77 | 0.80 |
| Lab/Path/1000 | 0.83 | 0.80 |

### Rx Utilization Correlation to the Rx Actual PMPMs: Non-Risk Adjusted

There should be strong correlations between the Rx rates to the Actual PMPMs and Actual TCRRVs as the only factor that is not accounted for in the Rx count metric is the intensity of the drug prescribed. The intensity includes generic usage as well as the variation in cost between drugs.

| **Rx** | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **Rx Actual PMPMs** | **Rx Actual TCRRVs** |
| Rx Count | 0.95 | 0.96 |

## *Risk Adjusted Correlations*

### Correlation Between the Risk Adjusted Place of Service Metrics and TCI and RUI

There should be some correlation between the high cost and resource intensive places of service and utilization measures to the TCI and RUI measures. The low intensive places of service and utilization should have a lower correlation to the overall TCI and RUI measures.

- The TCI is influenced by each provider group's overall cost per unit therefore there should be less correlation to the utilization metrics than the RUI. The following analysis will concentrate on the RUI.
- There is a high correlation between IP RUI and admit rate to the overall RUI.
- The professional RUI has a strong correlation with the overall RUI, while the E&M visits and lab/path services have a low correlation. This would indicate that the remaining professional services have a strong correlation to overall RUI.
- As expected there is no correlation between the Rx TCI and overall RUI as the ACG risk adjustment accounts for the variations in pharmacy usage.
- Both the standard and high tech radiology have some correlation to the RUI.

### Correlation Between Risk Adjusted Place of Service Utilization Metrics and Corresponding TCI and RUI

There should be a correlation between the place of service utilization metrics and the risk adjusted PMPMs and TCRRVs of the corresponding place of service. The magnitude of the correlation is dependent upon the utilization metric's penetration within the place of service and the cost and/or resource intensity of the metric. Since the risk adjustment accounts for variations in illness burden these correlations will be different from their non risk adjusted results displayed in the Correlations Overview section.

## Inpatient Utilization Metric Correlation to the Inpatient RUI – RiskAdjusted

There should be strong correlations between the risk adjusted admit rate and the inpatient TCI and RUI. The only two factors not measured by the risk adjusted admit rate are the intensity and price of the services performed.

- There is a high correlation between the risk adjusted admit rate and the inpatient TCI and RUI. This would indicate that the higher the risk adjusted admit rate the more likely a provider will have a higher than average TCI and RUI.

## Outpatient Utilization Metrics Correlations to the Outpatient TCI and RUI – Risk Adjusted

| **Outpatient** | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **OP TCI** | **OP RUI** |
| ER Cnt | 0.89 | 0.84 |
| OP Surgery | 0.29 | 0.39 |

- There is a high correlation between the risk adjusted ER count and the outpatient TCI and RUI. This would indicate that the higher the risk adjusted ER counts the more likely a provider will have a higher than average outpatient TCI and RUI.
- Outpatient surgery having less of a correlation to the outpatient RUI is an indication that these services are not the driving force behind the outpatient RUI performance.

## Professional Utilization Metrics Correlations to the Professional TCI and RUI – Risk Adjusted

| **Professional** | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **Prof TCI** | **Prof RUI** |
| E&M Visits | 0.41 | 0.46 |
| Lab/Path | 0.57 | 0.37 |

- The professional utilization metrics are moderately correlated to the professional TCI and RUI.
- This result is not unexpected as the professional place of service includes a significant amount of services beyond these two utilization measures (other professional services = 54%).
- It is also not unexpected as having higher than average utilization on diagnostic or management based services does not necessarily indicate a higher resource consuming patient.

## Rx Utilization Metric Correlation to the Rx TCI – RiskAdjusted

| **Rx** | **Correlatio n** |
|---|---|
| **Metric** | **RX TCI** |
| Rx Count | 0.73 |

- This indicates that more prescriptions equate to a higher Rx TCI, however there is no correlation between the Rx TCI and the overall RUI.

## *Detailed Provider to Provider Analysis*

The Total Cost of Care and Resource Use measure are designed to identify variations between providers accurately and correctly. This section of the analysis will compare findings and results from known utilization metrics to the findings and results from the Total Cost of Care and Resource Use measures. If there are differences in conclusions drawn, the analysis identifies the causes and determines which measure, utilization or Total Cost of Care and Resource Use is more accurate/correct.

Since each utilization metric is designed to measure a portion of health care services, a composite utilization measure is necessary to aide in the evaluation of the accuracy and correctness of the Resource Use measure. Since the TCI includes a cost per unit (price) component, the evaluation is more comparable between the RUI and utilization.

Composite Utilization: A utilization metric was created by weighting each of the underlying utilization metrics by the place of service percent of resources it represents of the total resources.

Composite Utilization Metric =

Inpatient (Admit Rate x 16%) +
Outpatient (average(ER rate, OP Surg Rate, High Tech Rad Rate) x 20%) +
Professional (average (E&M rate, Lab/Path Rate, Std Rad) x 45%) +
Pharmacy (Rx rate x 19%)

## Primary Care Provider Network Overview

### RUI vs. Risk Adjusted Composite Utilization Index

2009 Commercial, Continuously Enrolled, Excluding Babies and 65+
Indexed to the Metro Average

It is expected that the resources should correlated to the composite utilization metric.

| Provider Group | RUI | Admit | ER Count | OP Surgery | Hightech Rad | E&M | Lab/Path | Std. Rad | Rx Cnt | Composite Utilization |
|---|---|---|---|---|---|---|---|---|---|---|
| Provider O | 0.91 | 0.93 | 0.86 | 0.80 | 0.88 | 0.95 | 0.91 | 0.87 | 1.02 | 0.92 |
| Provider G | 0.93 | 0.51 | 0.82 | 0.96 | 1.02 | 1.05 | 1.08 | 0.90 | 1.07 | 0.93 |
| Provider D | 0.95 | 0.77 | 0.75 | 1.08 | 0.88 | 1.03 | 0.89 | 0.95 | 0.86 | 0.90 |
| Provider I | 0.97 | 0.99 | 0.91 | 0.94 | 0.93 | 0.98 | 1.07 | 1.00 | 0.94 | 0.98 |
| Provider P | 0.99 | 0.97 | 0.95 | 1.14 | 1.06 | 1.03 | 1.10 | 0.94 | 0.95 | 1.01 |
| Provider Q | 1.00 | 1.00 | 1.27 | 1.12 | 0.98 | 1.01 | 0.92 | 0.85 | 1.03 | 1.00 |
| Provider K | 1.00 | 1.19 | 1.23 | 1.17 | 1.07 | 1.00 | 0.86 | 0.85 | 0.93 | 1.00 |
| Provider E | 1.00 | 1.01 | 1.17 | 1.01 | 0.96 | 0.99 | 0.95 | 1.08 | 1.04 | 1.02 |
| Provider S | 1.01 | 1.03 | 1.04 | 1.13 | 1.14 | 0.96 | 0.78 | 1.04 | 1.01 | 0.99 |
| Provider F | 1.01 | 0.92 | 0.87 | 1.00 | 0.97 | 1.02 | 0.87 | 0.85 | 1.03 | 0.94 |
| Provider J | 1.01 | 0.78 | 1.45 | 0.98 | 0.80 | 0.98 | 0.95 | 0.97 | 1.36 | 1.03 |
| Provider M | 1.01 | 1.05 | 0.82 | 1.05 | 0.95 | 0.98 | 0.92 | 1.03 | 1.04 | 0.99 |
| Provider L | 1.02 | 1.05 | 0.89 | 0.86 | 1.41 | 1.04 | 1.09 | 1.11 | 1.00 | 1.05 |
| Provider A | 1.02 | 1.03 | 1.10 | 0.97 | 1.03 | 1.01 | 0.93 | 1.01 | 1.05 | 1.01 |
| Provider N | 1.03 | 0.97 | 0.90 | 1.03 | 1.03 | 0.99 | 1.00 | 0.93 | 1.08 | 1.00 |
| Provider B | 1.04 | 1.01 | 0.58 | 1.09 | 1.01 | 1.07 | 0.98 | 1.14 | 1.04 | 1.02 |
| Provider C | 1.06 | 1.16 | 1.42 | 0.99 | 1.09 | 1.01 | 0.94 | 1.22 | 0.93 | 1.07 |
| Provider H | 1.06 | 0.98 | 1.10 | 1.09 | 1.15 | 1.02 | 0.95 | 1.14 | 1.14 | 1.06 |
| Provider R | 1.07 | 1.10 | 0.75 | 1.02 | 0.94 | 0.98 | 1.07 | 0.94 | 0.97 | 0.99 |
| **Metro Total** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |

- The composite utilization index correlation to overall RUI is 0.77
- The composite index and the RUI have relatively the same index with the exception of provider groups F and R.
  - Provider F's composite utilization metric is 0.94 while their overall RUI is 1.01. The lower than average composite utilization metric is due to the significantly lower than average admit rate, ER services, lab/path and standard radiology services.
    - The professional services are being undervalued due to intensity as the professional RUI is 5% above average (see S12_Sample Score Report).

  - Provider R's slightly lower than average composite utilization metric is due to the lower than average ER visits, high-tech radiology, E&M, standard radiology services, and Rx count.
    - The weight of the admit rate in the composite score is undervalued due to intensity as the 10% higher than average admit rate translates to 24% higher than average inpatient resource use (see S12_Sample Score Report).

## High to Low Provider Contrast Analysis

The TCI and RUI should clearly identify providers that are high or low performing and be supported by the risk adjusted utilization metrics.

## Profile of a High Performing Provider (Low TCI and RUI)

- The four top performing providers achieve lower than average resource use with some common markers:
  - Lower than average admit and ER indices.
  - Standard radiology is at or lower than average for all providers.
  - E&M visits are within 5 points of average.
  - All other utilization markers do not have a clear direction.
- The place of service resource use index is near or below average for inpatient, outpatient, and professional components (see S12_Sample Score Report). The Rx TCI is high for one provider, however that provider has extremely low admits, which offset the Rx usage.

## Profile of a Low Performing Provider (High TCI and RUI)

- The lowest performing four providers have some common utilization markers which supports their higher than average resource use:
  - Higher than average in admits or ER or both.
  - These providers have a minimum of one of the other high resource intensive utilization metrics above average.
  - High tech and standard radiology is above average for all but one of the low performing providers. This one exception provider has 10% higher inpatient admissions.
  - E&M visits are relatively around average (one provider is at 1.07).
  - All other utilization markers do not have a clear direction.
- The place of service resource use index is above average for the professional component and at least one of the other 3 components (see S12_Sample Score Report).

## Profile of Providers that do not Fit the Peer Grouping (Excluded from Metro Primary Care Network)

The Total Cost of Care and the Resource Use measures are designed to evaluate providers that are similar in nature and are within the same peer group. Providers that have a significantly different patient mix or patient profile will stand out as outliers.

| | Average ACG Score | | | TCI | | | Resource Use Index | | |
|---|---|---|---|---|---|---|---|---|---|
| Provider Group | 2007 | 2008 | 2009 | 2007 | 2008 | 2009 | 2007 | 2008 | 2009 |
| Provider Y | 1.29 | 1.35 | 1.25 | 1.58 | 1.56 | 1.44 | 1.54 | 1.51 | 1.42 |
| Provider Z | 1.32 | 1.14 | 1.21 | 2.11 | 2.26 | 2.03 | 1.47 | 1.52 | 1.40 |

| Provider Group | RUI | Admit | ER Count | OP Surgery | Hightech Rad | E&M | Lab/Path | Std. Rad | Rx Cnt |
|---|---|---|---|---|---|---|---|---|---|
| Provider Y | 1.42 | 1.29 | 1.20 | 1.08 | 2.88 | 1.19 | 1.54 | 1.65 | 1.10 |
| Provider Z | 1.40 | 1.50 | 1.67 | 1.48 | 1.90 | 1.17 | 1.90 | 1.44 | 0.99 |

- Providers Y and Z have significantly higher ACG scores, which in and of itself is not an indication of an outlier provider.
- They also have significantly higher TCIs driven by 40% higher resource use.
- It is known that these providers treat patients that are high users and in need of a complex level of treatment.
- All of the utilization metrics are above average (except for Provider Z's RX count of 0.99)

## Detailed Provider to Provider Analysis – Selected Place of Service

The inpatient admit and the Rx count rates are highly correlated to their place of service RUIs as they encompass the majority of the services within the place of service and the only factors not measured is the unit cost and intensity of service.

## Expanded Inpatient Resource Use vs. Admit Rate Provider Analysis – Risk Adjusted

There is a strong correlation between the risk adjusted admit rate and the risk adjusted inpatient RUI (0.92, see Inpatient Utilization Metric Correlation to the Inpatient RUI section). The only 2 factors not measured by the risk adjusted admit rate are the intensity and price of the services performed. The RUI will account for the intensity of services performed.

- 9 out of 19 groups have lower than average IP RUI
- 1 out of 9 groups had slightly higher than average IP admissions, due to Provider B having a lower than average intensity.
- 10 out of 19 groups have higher than average IP RUI.
- 1 out of 10 groups had a slightly lower than average IP admissions, due to Provider I having more intensive than average admissions.

## Expanded Rx Total Cost Index vs. Rx Count Provider Analysis – RiskAdjusted

There is a strong correlation between the risk adjusted Rx count and the risk adjusted Rx TCI (0.73, see Rx Utilization Metric Correlation to the Rx TCI section). The only factor not measured by the risk adjusted Rx count is intensity (cost per unit is neutral for all providers). Variations in costs of pharmaceuticals and generic rates would express themselves through intensity and be accounted for in the Rx TCI, but not the Rx count metric.

- 9 out of 19 groups have lower than average Rx TCI.
- 4 out of 9 groups have slightly higher than average Rx fills.
- Providers M and O have higher than average percent generic rate which influences the Rx TCI.
- Providers Q and S have slightly lower than average percent generic rate. The higher than average Rx count and lower than average Rx TCI is due to the prescriptions being less resource intensive/costly than average.
- 10 out of 19 groups have higher than average Rx TCI.
- 2 out of 10 groups have lower than average Rx fills.
- Provider R's percent generic rate is 69% compared to the metro average of 74%, which drove their higher than average Rx TCI.
- Provider P had a slightly lower than average percent generic rate. The lower than average Rx count and average Rx TCI is due to the prescriptions being more resource intensive/costly than average.

## Definitions

### Service Category

- Inpatient: Claims on a 1450 claims form and one of the following criteria
  - Room and Board Revenue codes: 100-189, 200-219, 650, 655, 1000-1005
  - Bill Type code: 21, 28, 66, 86
  - Bill Type code of 11 and a revenue code of 190
- Outpatient all other 1450 claim forms
- Professional all 1500 claim forms
- Rx – All pharmacy data

## Total Cost of Care Validity Metric Overview

### Utilization Metrics

Admits

An inpatient admission.

ER Count

An outpatient claim that includes at least one revenue code between 450-

459. E&M Count

E&M CPT codes from a professional claim.

Lab\Path

All Laboratory and Pathology CPT codes.

Standard Radiology

All radiology CPT codes that are not considered high technology radiology (MRI, CT, nuclear medicine, PET).

Outpatient Surgery

All outpatient visits that include one surgical CPT.

High Technology Rad

CPT codes from the professional or outpatient place of service that are considered an MRI, CT, nuclear medicine or PET scan. Only one bill is counted if two are submitted for one patient.

Rx Count

Script count.

Percent Generic

The percent of prescription that are generic.

## Other Metrics

Actual PMPM

The actual spend divided by the member months of the population. These are non risk adjusted numbers.

ACG Score

At any given level it is the sum of a (member's assigned ACG cell weight x their member months divided by the total member months) of the given level (aka Average ACG weight at any given level).

TCRRV

Total Care Relative Resource Value – is a price neutral value that is relative within and across all places of service and types of treatment. In essence it is a standard fee schedule of all services within the health care continuum.

TCRRV PMPM

The actual TCRRVs divided by the member months of the population. These are non risk adjusted numbers.

TCI

Total Cost Index – the ACG risk adjusted spend PMPM divided by the analysis population's ACG adjusted spend PMPM.

RUI

Resource Use Index – the ACG risk adjusted TCRRV PMPM divided by the analysis population's ACG adjusted TCRRV PMPM.

NATIONAL QUALITY FORUM

# COST AND RESOURCE USE MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

**To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return**

## Brief Measure Information

**NQF #:** 1604

**Measure Title:** Total Cost of Care Population-based PMPM Index

**Measure Steward:** HealthPartners

**Brief Description of Measure:** Total Cost of Care reflects a mix of complicated factors such as patient illness burden, service utilization and negotiated prices. Total Cost Index (TCI) is a measure of a primary care provider's risk adjusted cost effectiveness at managing the population they care for. TCI includes all costs associated with treating members including professional, facility inpatient and outpatient, pharmacy, lab, radiology, ancillary and behavioral health services.

A Total Cost of Care Index when viewed together with HealthPartners (NQF-endorsed #1598)Total Resource Use measure provides a more complete picture of population based drivers of health care costs.

**Developer Rationale:** By measuring population based total cost of care, health plans and providers can improve the affordability of health care without sacrificing quality. HealthPartners' TCI gives provider groups valuable information on the cost of care and, when viewed in conjunction with resource use and quality metrics, information on the efficiency of care. The HealthPartners TCI measure is a population-based, patient-centered, total cost of care measure that crosses all categories of health services. This is in contrast to the many, episodic based measures available in the market today. Both population based and episodic based measures are important and complimentary but a key benefit of population based measures is helping to better understand potential overuse & underuse (e.g., although efficient at spine surgery, may be performing too many).

**Resource Use Measure Type:** Per capita (population- or patient-based)

**Data Source:** Claims (Only)

**Level of Analysis:** Clinician : Group/Practice, Population : Community, County or City

**Costing Method:** Actual prices paid

The Total Cost of Care considers 100% of health care services in the Total Cost Index and is calculated on a risk-adjusted paid per member per month basis as well benchmarked to a peer group. The paid amount (i.e., allowed) is inclusive of both plan and member liability.

**Tested Population:** The validity and reliability testing of the measures was conducted with the HealthPartners' commercial population which is 470,000 members. For purposes of testing income disparities for the SES analysis, Medicaid was included in addition to commercial which is the combined total membership of 530,000 members.

**Resource Use Service Categories:**

**Attribution Approach**

The level of analysis for this measure could be an entire health plan, provider group, employer group and/or geographic in nature. Measure was tested using commonly used Attribution Algorithm in an open access market (plurality model, using most recent visit as a tie breaker):

- Include twelve months based on first date of service for the measurement year (e.g. January 1 – December 31) of professional claims experience, with three months of paid claims run out to allow for claims lag.
- Exclude all services that are not office based
- Exclude convenience care clinic visits and hospice services
- Exclude a providers that are not a physician, physician assistant or nurse practitioner
- Assign each service line a specialty based on the servicing physician's practicing specialty or credential specialty if practicing specialty is not available.
- Include only the following specialties:

- Family Medicine, Internal Medicine, Pediatrics, Geriatrics, OB/GYN

**IF Endorsement Maintenance – Original Endorsement Date:** Jan 31, 2012 **Most Recent Endorsement Date:** Jan 31, 2012

# Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measure still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

## Criteria 1: Importance to Measure and Report

### 1a. High Priority

**1a. High Priority**. This requirement involves demonstrating that the measure focus addresses one of the following:

- A specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF.
- A demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality).

**Summary of information provided to fulfill the High Priority requirement**

- To demonstrate the importance of measuring cost, the developers cite data demonstrating healthcare spending constitutes a high proportion (17%) of the United States gross domestic product (GDP) and high healthcare costs contributes to adults forgoing healthcare.
- The developers suggest that this measure can support a comprehensive measurement system to identify areas of overuse.

**Preliminary rating for High Priority:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

### 1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
**Maintenance measures – increased emphasis on gap and variation**

**1b. Performance Gap.** This requirement involves demonstrating a resource use or cost problems exist and there is an opportunity for improvement (i.e., data demonstrating variation in the delivery of care across providers and/or population group (disparities in care)).

- The developer presents performance data from 2015 dates of service from the multi-stakeholder community collaborative, Minnesota Community Measurement (MNCM) measured the Total Resource Use of 257 provider groups, representing 1.5 million patients receiving care. The 2015 risk-adjusted total cost of care per member per month on average was $474, with a range of $365 to $916. Eighty percent of provider groups were between $394 and $555 per member per month. The developer did not provide data on changes in performance over time.
- It is unclear if the performance gap demonstrated is based on the measure as specified.
  - *The developer has clarified that these analysis used the measures as specified.*

**Disparities**

- To examine differences in measure scores by age and gender, the developer examined the distribution of scores in single specialty obstetric and pediatric groups. Data from these analyses were not provided, but the developer states scores were uniformly distributed and not clustered.

***Questions for the Committee:***

- *Is there a gap in care that warrants a national performance measure?*

**Preliminary rating for opportunity for improvement:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

### 1c. Measure Intent

**1c. Intent of the resource use measure.** This requirement involves describing the measure intent of the resource use measure and the measure construct.

- The intent of this measure is to allow measure implementers to better understand and measure overuse and underuse to drive person-centered management and accountability.

- A population-based measure complements conditions and episode-based measure for a complete view of utilization across the measurement year.

***Questions for the Committee:***

- *Is the measure clearly described?*
- *Is it appropriate to measure costs in this way? At this level of analysis?*
- *Are the costs included appropriate and consistent with the measure intent?*
- *Is there at least one thing that providers can do to achieve a change in the measure results?*

**Preliminary rating for measure intent:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)**

***1a. High Priority***

Comments:

**Very important topic area.   The combination of the TCI with the RUI can provide very valuable insights for provider groups.
**Cost is definitely high impact, so the topic meets High Priority definition
**I agree with high priority.
**Total cost measures for primary care are important assessments of crucial domain within the IOM Quality model. Using total payments per patient for a limited number of commercial payers may provide a limited perspective on the true "cost behavior" of providers, particularly with increasing use of limited networks or special contracts between those payers and selected providers.
**Yes, measure meets sub-criterion. NQF assessment captured adequately.
**Not sure if high priority.
**High--a priority for CMS/Medicare/HHS

***1b. Performance Gap***

Comments:

**Clear variation in total cost across provider groups.
**I am less confident that this metric is providing evidence of a gap in care.  The Developers note that the measure is being widely used, but offer only Minnesota and Wisconsin as examples.  Shouldn't there be greater data available for a maintenance measure?
**It is plausible to believe that the variance in PMPM cost in the examples cited exists in other populations and health systems.
**Variability is seen across primary care clinics indicating potential opportunities for improvement.
**Yes, measure meets sub-criterion. The 2015 risk-adjusted total cost of care per member per month on average was $474, with a range of $365 to $916. Eighty percent of provider groups were between $394 and $555 per member per month. Other evidence also shows significant variation and thus room for improvement, e.g. Dartmouth Atlas. Uniform distribution for age and gender for some practice groups.
**The measure shows substantial variation across provider groups, with a substantial variation in prices paid contributing to this variation.  No analysis of disparities in care.
**Moderate---the gap depends on the unit of analysis.  So will vary depending on how measure is used and within that specific population. Intent is clear.

**Criteria 2: Scientific Acceptability of Measure Properties**

**2a. Reliability**

**2a1. Reliability  Specifications**

**Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures**

**2a1. Specifications.** This requirement involves providing the full specifications for the measure so that it can be implemented consistently within and across organizations and allow for comparability. Electronic health record (EHR) measure specifications are based on the quality data model (QDM).

**Data source(s):**

- Claims (measure and risk adjustment model)

**Specifications:**

- This per capita (population- or patient-based) measure calculates the total cost of care of a commercial population and is expressed as a ratio.
- To interpret, a score greater than 1.00 indicates a higher paid risk adjusted PMPM value, compared to a peer group average; a score less than 1.00 indicates less paid risk adjusted PMPM value, compared to a peer group average.
  - The choice of a peer group is at the discretion of the measure user and can include the internal medicine, family medicine, pediatrics, geriatrics, and OB/GYN specialties and physician, physician assistant, and nurse practitioner provider types  The peer group's average is set at the benchmark.
- The numerator is calculated as the sum of (Total Medical Cost / Medical Member Months) + (Total Pharmacy Cost / Pharmacy Member Months).
- The Johns Hopkins Adjusted Clinical Grouper (ACG) risk scores constitutes the measure's denominator.
- The developer provides the following steps regarding the measure's construction logic:
  - Obtain all claims that have a date of service in the measurement year. The measurement year is not explicitly defined by the developer, but they provide an example year as running from January 1st to December 31st.
  - Include members enrolled for a minimum of 9 months in the measurement year
  - Include commercial population only
  - Attribution - the developer acknowledges the attribution approach used by measure implementers may vary according the implementer's business purposes and unit of measurement, but does provide the following attribution guidelines:
    - Include twelve months based on first date of service for the measurement year (e.g. January 1 – December 31) of professional claims experience, with three months of paid claims run out to allow for claims lag.
    - Exclude all services that are not office based
    - Exclude convenience care clinic visits and hospice services
    - Exclude a providers that are not a physician, physician assistant or nurse practitioner
    - Assign each service line a specialty based on the servicing physician's practicing specialty or credential specialty if practicing specialty is not available.
    - Include only the following specialties:
      - Family Medicine, Internal Medicine, Pediatrics, Geriatrics, OB/GYN
  - Costing Method – actual prices paid are used for this measure and 100% of healthcare services are included. The amount paid is inclusive of both plan and member liability.
- Missing data
  - For members that have their pharmacy benefits carved-out, a proxy of  the provider's risk-adjusted pharmacy costs is included. This allows for a calculation of total PMPM .
  - For additional carve outs, the developer indicates the "lowest common denominator principle" should be applied, meaning all services carved out of one segment of input data should be carved out of the measure for all segments of input data and all input components (e.g., PMPMs, attribution, and risk adjustment).
- Clinical logic
  - The developer states clinical logic is not applicable given this is a population-based measure that applies to all care settings and conditions.
- Adjustments for comparability: The measure developer used the following exclusion criteria and risk adjustment approach. The developer does not include explicit inclusion criteria in the measure submission.
  - Exclusion criteria:
    - Members over age 64
    - Members under age 1
    - Member enrollment less than nine months during the one year measurement time window

        - Members who are not attribute to a primary care providers
        - Dollars per member above $125,000
    - Risk Adjustment
        - The measure is risk adjusted for age, gender, and diagnosis using the Adjusted Clinical Group (ACG) method.
        - The ACG method involves:
            - Grouping International Classification Diagnosis (ICD) diagnosis codes into 32 diagnosis groups (i.e., Aggregated Diagnosis Groups (ADGs)). These ADGs are clinically similar and expected to have similar need for healthcare resources.
            - Adjusted Clinical Groups (ACGs) are created from the ADG assignments and are defined by morbidity, age, and sex. Individual members are then assigned to a single ACG category, which quantifies their risk.

**Changes to specifications since previous evaluation:**

- The developer reported one change to the measure specifications. Previously, members were if their total medical and pharmacy costs exceeded $100,000. The developers increased this amount to $125,000 to account for the natural rise in healthcare costs over the past several years.

***Questions for the Committee***:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*
- *Is the clinical logic clear? Is the construction logic clear?*

## 2a2. Reliability Testing

**Testing attachment**

**Maintenance measures – less emphasis if no new testing data provided**

**2a2. Reliability testing:** This requirement involves demonstrating that the measure data elements are repeatable, produce the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

**For maintenance measures, summarize the reliability testing from the prior review:**

- In the 2012 submission (see Appendix A), the developer provided a summary of the measure score reliability testing conducted using data obtained HealthPartners' primary care Twin Cities metro area providers for the calendar years of 2007, 2008, and 2009. The testing sample included 19 individuals providers and 268,912 (2007), 272,491 (2008), and 303,639 (2009) members. The methods for assessing measure score reliability included – a 90% random sample, a bootstrapping technique, and analysis of performance overtime. In the 90% and bootstrapping methods, reliability was measure as the mean of the variance between the sampling results and the actual results. Results from each method are summarized below:
    - 90% Sample – Variance ranging from -0.0069 to 0.00083 in 2009
    - Bootstrapping – Variance ranging from -0.00067 to 0.00252 in 2009
    - Provider performance - Relatively consistent across all three years with an average difference of 0.031
- In the 2012 review, the Committee found the testing adequately demonstrated the measure's reliability and passed the measure on the reliability criterion (High-8; Moderate-6; Low-4; Insufficient-0; NA-0).

**Describe any updates to testing:**

- For this maintenance submission, validity and reliability testing of the measures was conducted with HealthPartners' commercial population which is 470,000 members. (see testing details below).

**SUMMARY OF TESTING**

**Reliability testing level** ☒ **Measure score** ☐ **Data element** ☐ **Both**

**Reliability testing performed with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**

**Method(s) of reliability testing**
Updated Testing

- To demonstrate measure score reliability, the developer conducted the following analyses:
  1. Comparing actual measure scores to scores calculated by two sampling methods:
     - Bootstrapping
     - A 90% random sample

**Results of reliability testing**
Updated Testing

- The results of the reliability testing are summarized below:

| Testing Method | Results<br>Difference between Actual Score & Sampling Scores | Results<br>Variation |
|---|---|---|
| Bootstrapping | Range: -0.0059 to 0.0075 | Within groups <1%; Between groups >110% |
| 90% Sample | Range: -0.0022 to 0.0012 | N/A |

***Questions for the Committee:***

o *Is the test sample adequate to generalize for widespread implementation?*

o *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

**Guidance from the Reliability Algorithm Precise specifications (Box 1) → Empiric reliability testing (Box 2) → Score-level testing (Box 4) → Appropriate method (Box 5) → High certainty that measure results are reliable (Box 6a)**
**Preliminary rating for reliability: ☒ High ☐ Moderate ☐ Low ☐ Insufficient**

**2b. Validity**
**Maintenance measures – less emphasis if no new testing data provided**

**2b1. Validity: Specifications**

**2b1. Validity Specifications:** This requirement involves demonstrating that the measure specifications are consistent with the measure intent described under criterion 1c and capture the most inclusive target population.

**Specifications consistent with intent described in 1c. ☒ Yes ☐ Somewhat ☐ No**

***Question for the Committee:***

o *Does the Committee agree the specifications are consistent with the intent of the measure?*

o *Is the attribution approach consistent with the measure intent?*

o *Does the accountable entity have reasonable control over the resources measured?*

**2b2. Validity testing**

**2b2. Validity Testing** This requirement involves demonstrating that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided.
**For maintenance measures, summarize the validity testing from the prior review:**

- In the 2012 submission (see Appendix A), the developer provided a summary of construct validity testing conducted using data obtained HealthPartners' primary care Twin Cities metro area providers for the calendar years of 2007, 2008, and 2009. The testing sample included 19 individuals providers and over 300,000 members in the 2009 sample. Construct validity was tested by examining the correlations between the measure score and known utilization metrics and ACG scores.

- In the 2012 review, the Committee passed the measure on validity testing (validity testing: High-7; Moderate-5; Low-5; Insufficient-0; NA-0); however, the Committee expressed concerns about the developer's attribution guidelines. The Committee was concerned that while attribution is based on outpatient resource use, the measure specifications include inpatient costs, which could result in a provider being held responsible for an individual's inpatient visit before ever seeing the patient in an outpatient visit. The Committee expressed concern that this might dis-incentivize providers to take patients who have not seen a primary care provider. There was concern with respect to the level of analysis and the need for clarity around how a physician group was defined. The developer defined a physician group as 2 or more physicians, with a recommended minimum of 600 patients in the sample.

**Describe any updates to validity testing:**

- For this maintenance submission, the developer summarized updated validity testing conducted using provider data from 2014 and 2015. The validity and reliability testing of the measures was conducted with HealthPartners' commercial population which is 470,000 members. This updated validity testing consisted of correlations the measure components (i.e., ACG scores, unadjusted costs) and measure score with other markers of utilization.

**SUMMARY OF TESTING**

**Validity testing level** ☐ **Measure score** ☐ **Data element testing against a gold standard** ☒ **Both**

**Method of validity testing of the measure score:**

☐ **Face validity only**

☒ **Empirical validity testing of the measure score**

**Validity testing method:**

- *Data element validity*
  - To demonstrate data element validity, the developer conducted a series of correlation analyses:
    - Measure components (i.e., ACG scores & Non-risk adjusted per member per month value (Non-Risk Adjusted PMPMs))
      - ACG Risk-adjusted Total Cost Index (i.e., the measure score)
      - ACG risk-adjusted Resource Use Index (RUI) (i.e., measure 1598)
      - Non-risk adjusted Total Cost Relative Resource Values (TCRRVs)
      - Price
    - Measure component - Non-Risk Adjusted PMPMs with non-risk adjusted rates of utilization:
      - Inpatient Admits per 1,000
      - ER per 1,000
      - Outpatient surgery per 1,000
      - High Tech Radiology per 1,000
      - E&Ms per 1,000
      - Lab/Path per 1,000
      - Standard radiology per 1,000
      - Pharmacy per 1,000
    - Measure Components with Composite Utilization
- *Measure score validity – Empirical Testing*
  - To demonstrate measure score validity, the developer conducted a series of correlation analyses:
    - ACG Risk-adjusted Total Cost Index (i.e., the measure score) with:
      - Hospital based Total Cost of Care Index
      - Professional Total Cost of Care Index
      - Pharmacy Total Cost of Care Index
      - ACG risk-adjusted Resource Use Index (RUI) (i.e., measure 1598)
      - Total Price
    - Service Category TCI (i.e., Inpatient, Outpatient, Professional, Pharmacy) with risk-adjusted service category metrics:
      - Inpatient admit rate
      - ER count
      - Outpatient surgery

        - High tech Radiation
        - E&M Visits
        - Lab/Path
        - Standard Radiology
        - Prescription (Rx) Count
    - Measure Score with Composite Utilization
    - Measure Score Over time

- *Measure score validity – Face Validity*
    - To demonstrate measure score face validity, the developer cites their process of sharing measure scores and measure methodology with measured providers.
    - NQF requires a systematic assessment of face validity to be assessed. A systematic assessment of face validity is used when a panel of experts evaluates the measure specifications and measure testing to assess if the measure is an accurate reflection of performance. Results from a panel of experts is not included.
    - *Additional face validity information provided by the developer:*
        - *HealthPartners measures have been systematically evaluated for face validity by the following organizations, each convening panels of experts:*
        - *HealthPartners: Internally reviewed by Cost Assessment Committee (medical directors, network management, health informatics). Since 2010, transparent quarterly reporting to 60+ provider groups in the HealthPartners network. All providers have 45 days to review prior to public reporting.*
        - *Minnesota Community Measurement: Reviewed by two multi-stakeholder groups - Cost Technical Advisory Group (TAG) – including patients, providers and purchasers, and the Measurement and Reporting Committee (MARC) - consumers, providers, health plans, purchasers prior to public reporting*
        - *Total Cost of Care – measure used as specified (pages 1-5): http://mncm.org/wp-content/uploads/2013/04/2014.11.12-MARC-Minutes_Approved.pdf*
        - *Total Resource Use – measure used as specified, known as 'RRU' in this document, (pages 2-3): http://mncm.org/wp-content/uploads/2016/11/2016.09.14-MARC-Minutes_Approved.pdf*
        - *Network for Regional Healthcare Improvement (NRHI) – The RWJF grant funded to produce and distribute practice level regional Total Cost of Care Reports. The first phase represented five regional health care improvement collaboratives in Colorado, Maine, Missouri, Minnesota and Oregon. Each region produced and distributed practice level reports in their communities, and a benchmark approach was developed and tested, and have committees and board of directors that oversee the work.*

**Validity testing results:** (highlighted values are those directly relevant to the measure under evaluation)

- Data element validity testing results
    - Correlation between measure components, ACG Score and Non-Risk Adj PMPMs and other metrics

| | Correlation Coefficient | |
|---|---|---|
| **Metric** | **ACG** | **Non-Risk Adj PMPMs** |
| Non-Risk Adj PMPM | 0.62 | 1.00 |
| Non-Risk Adj TCRRVs | 0.88 | 0.78 |
| ACG Risk Adj TCI | 0.03 | 0.79 |
| ACG Risk Adj RUI | 0.14 | 0.45 |
| Price | -0.09 | 0.57 |

        - The developer notes there is a high correlation of the measure components to one another and each component's correlation with the Non-Risk Adj TCRRVs as sufficient evidence for the validity of the measure components.

        - The correlation between the non-risk adjusted PMPM and the ACG Risk Adjusted TCI is 0.79.
        - The developer attributes the low correlated between ACG and Price to fact that ACG is an estimate of expected resource use whereas price is the unit cost of services actually provided.

- Measure component - Non-Risk Adj PMPMs with non-risk adjusted rates of utilization:

| **Non-Risk Adjusted** | **Correlation Coefficient** | |
|---|---|---|
| *Service Category*<br>**Metric** | ***Non-Risk Adj Service Category PMPMs*** | ***Non-Risk Adj Service Category TCRRVs*** |
| ***Inpatient*** | | |
| Admits/1000 | 0.67 | 0.82 |
| ***Outpatient*** | | |
| ER/1000 | 0.67 | 0.52 |
| OP Surgery/1000 | 0.60 | 0.68 |
| HighTech Rad/1000 | 0.45 | 0.67 |
| ***Professional*** | | |
| E&M/1000 | 0.63 | 0.71 |
| Lab/Path/1000 | 0.77 | 0.83 |
| Std Rad/1000 | 0.49 | 0.72 |
| ***Pharmacy*** | | |
| Rx/1000 | 0.73 | 0.80 |

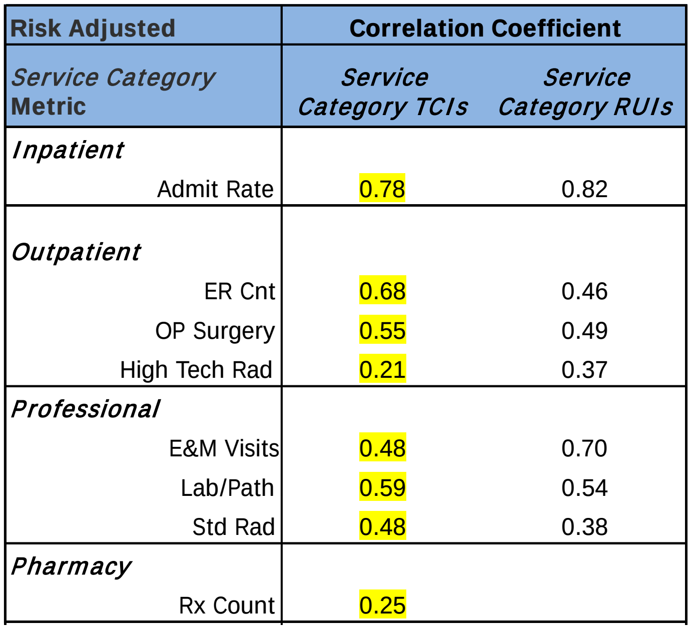
    - Measure Components with Composite Utilization

| **Non-Risk Adjusted** | **Correlation Coefficient** | | |
|---|---|---|---|
| **Metric** | **ACG** | **Non-Risk Adj PMPMs** | **Non-Risk Adj TCRRVs** |
| Composite Utilization | 0.74 | 0.69 | 0.87 |

- *Measure score validity – Empirical Testing*

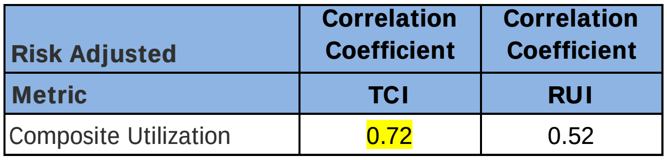
| **Risk Adjusted** | **Correlation Coefficient** | | |
|---|---|---|---|
| **Metric** | **TCI** | **RUI** | **Price** |
| Hospital TCI | 0.74 | | |
| Prof TCI | 0.73 | | |
| Rx TCI | 0.16 | | |
| Hospital RUI | | 0.30 | |
| Prof RUI | | 0.74 | |
| Total RUI | 0.39 | | |
| Hospital Price | | | 0.86 |
| Prof Price | | | 0.83 |
| Total Price | 0.87 | | |

- Service Category TCI (i.e., Inpatient, Outpatient, Professional, Pharmacy) with risk-adjusted service category metrics:



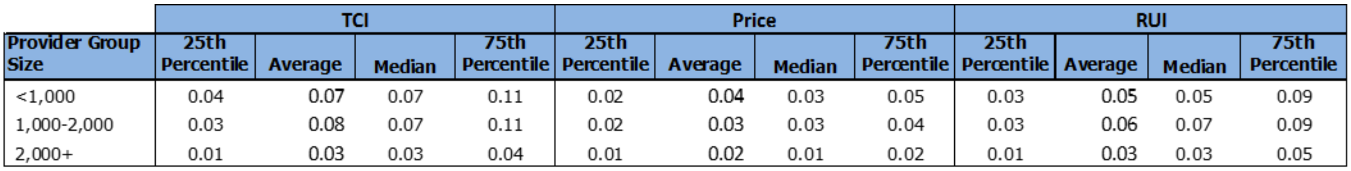
| Risk Adjusted | Correlation Coefficient | |
|---|---|---|
| *Service Category* Metric | *Service Category TCIs* | *Service Category RUIs* |
| ***Inpatient*** | | |
| Admit Rate | 0.78 | 0.82 |
| ***Outpatient*** | | |
| ER Cnt | 0.68 | 0.46 |
| OP Surgery | 0.55 | 0.49 |
| High Tech Rad | 0.21 | 0.37 |
| ***Professional*** | | |
| E&M Visits | 0.48 | 0.70 |
| Lab/Path | 0.59 | 0.54 |
| Std Rad | 0.48 | 0.38 |
| ***Pharmacy*** | | |
| Rx Count | 0.25 | |

- Measure Score with Composite Utilization



| Risk Adjusted | Correlation Coefficient | Correlation Coefficient |
|---|---|---|
| Metric | TCI | RUI |
| Composite Utilization | 0.72 | 0.52 |

- Measure Score over time



| | TCI | | | | Price | | | | RUI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Provider Group Size | 25th Percentile | Average | Median | 75th Percentile | 25th Percentile | Average | Median | 75th Percentile | 25th Percentile | Average | Median | 75th Percentile |
| <1,000 | 0.04 | 0.07 | 0.07 | 0.11 | 0.02 | 0.04 | 0.03 | 0.05 | 0.03 | 0.05 | 0.05 | 0.09 |
| 1,000-2,000 | 0.03 | 0.08 | 0.07 | 0.11 | 0.02 | 0.03 | 0.03 | 0.04 | 0.03 | 0.06 | 0.07 | 0.09 |
| 2,000+ | 0.01 | 0.03 | 0.03 | 0.04 | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 | 0.03 | 0.03 | 0.05 |

***Questions for the Committee:***

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *For data element validity and measure score validity, are the correlations in the expected direction and of the expected magnitude?*
- *Are the correlations between the measure score and place of service metrics sufficient for demonstrating measure score validity?*

## 2b3-2b7. Threats to Validity

**2b3. Exclusions**: This requirement involves demonstrating that the exclusions are:

- supported by the measure intent

AND/OR

- There is a rationale or analysis demonstrating that the measure results are sufficiently distorted due to the magnitude and/or frequency of the non-clinical exclusions;

AND

- Measure specifications for scoring include computing exclusions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion);

AND

- Patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

**Summarize approach and analysis of exclusions**

- Members with the following characteristics are excluded from the measure:
    - Members over age 64
    - Members under age 1
    - Member enrollment less than 9 months during the one year measurement time window
    - Members not attributed to a primary care provider
    - Dollars per member above $125,000 are excluded (i.e., truncated)
        - In the 2012 submission, the exclusion amount was $100,000
- To examine the determine the effect of the updated $125,000 level exclusion, multiple regression models examining various exclusion amounts were examined, specifically the percentage of patients excluded and the models' $R^2$ value, which is a measure how close the data are fitted to the regression line.
- The results from this analysis showed the percentage of patients excluded and the $R^2$ value are similar between the $100,000 exclusion level and the $125,000 exclusion level.

| Exclusion Level | % Members Excluded | % Dollars Included | $R^2$ Value |
|---|---|---|---|
| $100,000 | 0.5 | 92 | 0.473 |
| $125,000 | 0.3 | 94 | 0.472 |

- The developers states testing shows the exclusion of members under 1 and those without 9 months of enrollment during the measurement yet has little impact on the model's $R^2$ value, but do not provide specific data to support this claim.
- Analyses were not conducted examining the effect of excluding Members over 64, rather the developer state they are excluded due to potential incomplete claims data from Medicare eligible beneficiaries.

***Questions for the Committee:***

- *Are the exclusions consistent with the intent of the measure? Are carve-outs appropriately addressed?*
- *Are any patients or patient groups inappropriately excluded from the measure? Specific patient groups to consider include patients who died during the measurement period, patients who were transferred, and patients enrolled in Medicare Advantage plans.*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: This requirement involves specifying an evidence-based risk-adjustment strategy (e.g., risk models, risk-stratification) that is based on patient clinical factors that influence the measured outcome and are present at the start of care and has demonstrated adequate discrimination and calibration. If a risk adjustment strategy is not provided, a rationale or data to support no risk-adjustment/-stratification must be provided.

**Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

**Conceptual rationale for SDS factors included ?** ☒ **Yes** ☐ **No**

**SDS factors other than age and gender included in risk model?** ☐ **Yes** ☒ **No**

**Risk adjustment summary**

- The risk adjustment approach utilized in the measure is the Johns Hopkins Adjusted Clinical Grouper (ACG) method, which adjusts for age, gender, and diagnosis (i.e., clinical risk). A conceptual rationale for this risk adjustment approach is provided.

- The risk adjustment approach involves:
  - Grouping International Classification Diagnosis (ICD) diagnosis codes into 32 diagnosis groups (i.e., Aggregated Diagnosis Groups (ADGs)). These ADGs are clinically similar and expected to have similar need for healthcare resources.
  - Adjusted Clinical Groups (ACGs) are created from the ADG assignments and are defined by morbidity, age, and sex. Individual members are then assigned to a single ACG category, which quantifies their risk.
- Individual member ACG weights: Individuals are assigned to an ACG actuarial cell that has a corresponding weight reflecting relative illness burden. The ACG weight is then multiple by their number of eligible member months.
- Providers' ACG Scores are calculated as the sum of their attributed members ACG weights.
- Given the ACG risk adjustment approach is owned by Johns Hopkins, the developer does not provide a summary of statistical results of the analyses conducted on ACG risk model as that information is proprietary.

**Empirical Summary of SDS**

- Two measures of income - tract-level income, obtained from U.S. Census Tract data, and the household-level, obtained from a commercially licensed consumer database purchased by HealthPartners – were used to examine the impact of SDS on the measure scores.
- Two multiple linear regression equations were analyzed:
  1. Equation 1: Tract-level income, ACG risk score, and insurance product (i.e., Commercial vs Medicaid) were regressed on total reimbursed amount per member per month; and
  2. Equation 2: Household-level income, ACG risk score, and insurance product (i.e., Commercial vs Medicaid) were regressed on total reimbursed amount per member per month
- Results from both Census tract-level and household-level data sources show that income does not significantly impact the measure scores after risk adjusting for age, gender, and clinical risk, and stratifying by insurance type. The ACG score and the insurance type have a significant impact on the cost and resource use measures' variation and income has no discernible impact.

**Risk Model Discrimination and Calibration**

- For model discrimination, the developers provider the correlations of non-risk adjusted PMPM and ACG scores with other metrics of utilization. Discrimination and calibration statistics were not provided.

***Questions for the Committee:***

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors beyond age and gender in their risk-adjustment model?*

2b5. Meaningful difference: This requirement involves demonstrating, through data analysis, that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically meaningful differences in performance.

- To demonstrate the measure's ability to identify meaningful differences, the developer provides additional guidance on interpreting measures scores and a summary of performance in 66 providers groups.
- The developer states that statistically significant differences are not necessary as the measure is based on a full population and offers additional methods for examining differences including percentile, percent from the mean, and others. The choice of method would be dependent upon the business purpose.

***Question for the Committee:***

- *Does this measure identify meaningful differences about cost or resource use?*

2b6. Comparability of data sources/methods: This requirement involves demonstrating that if multiple data sources/methods are specified, they produce comparable results.

N/A

2b7. Missing Data: This requirement involves describing how missing data are handled and demonstrating that the presence of missing data does not bias the measure.

- The developer states that this is a full population-based measure and all data is included.
- For members that have their pharmacy benefits carved-out, a proxy of the provider's risk-adjusted pharmacy costs is included. This allows for a calculation of total PMPM .
- For additional carve outs, the developer indicates the "lowest common denominator principle" should be applied, meaning all services carved out of one segment of input data should be carved out of the measure for all segments of input data and all input components (e.g., PMPMs, attribution, and risk adjustment).

**Guidance from the Validity Algorithm** Precise specifications (Box 1) YES → Empirical testing conducted with measure as specified (Box 2) YES → Measure Score validity testing conducted (Box 6) YES → Testing method described and deemed appropriate (Box 7) YES → Moderate certainty that the measure score is a valid indicator of quality

**Preliminary rating for validity:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b)**

***2a1. & 2b1. Specifications***

***2a1. Reliability Specifications***

Comments:

**Elements are clearly defined.

**For sample size, they note that the testing was done with an attributed population of 600 members. Is this within a group? By a payer? By a plan? Not clear. Attribution was a big concern in 2012 when this was last reviewed. I am not sure that concern has been answered.

**I would like to know more about the risk adjustment method used.

**Is Attribution method part of the TCOC process? If attribution method differs from that proposed, what impact does that have on NQF approval? In Minnesota Community Measurement application of TCOC, problems with their attribution method result in the attribution of patients seen in specialty clinics (e.g., cancer center, GI clinic) by NP/PAs and medical residents prior to their board certifications as primary care patients because these groups are counted as internal medicine providers.

**Yes, measure meets sub-criterion. NQF assessment captured adequately.

**Reliability is high; the measure is clearer defined and implemented.

**High--clearly defined

***2a2. Reliability Testing***

Comments:

**Overall not concerned about reliability

**The reliability testing seems solid. I have some worries about risk adjustment that will be detailed below.

**The sample sizes are adequate, but I would like to see reliability testing in older (65+) populations and underserved populations.

**Although the population was of good size, it appears that all claims were from one payer and a restricted area of the west north central region of the the US. This part of the country has a larger portion of large group practices than much of the US. Generalizability would be strengthened by including more payers and a wider range of provider types.

**Yes. Additional reliability testing presented since last endorsement.
Conducted at measure score level.

**Ok.

**Moderate. I didn't see reliability estimates to understand whether measure distinguishes between providers (signal to noise). It is mentioned in text that they tested this but I don't see results.

***2b1. Validity Specifications***

Comments:

**Claims based measure with valid specifications

**Specifications seem reasonable. The score is easy to interpret.

**Using total payments sounds like a good idea, but growing use of restricted networks and higher payments for patient utilization outside networks may confound TCOC performance with organizational contract negotiations, thereby limiting provider control over their performance.

**There does not appear to be an inconsistency.
**Specifications are clear and consistent with evidence.  While there may be validity concerns, they are not in specifications.
**High--clearly defined

***2b2. Validity Testing***

Comments:
**Committee needs to walk through the validity testing and updated information in some detail.
**Was the measure tested both within and between different specialties?  It is perfectly appropriate to consider OB/GYN as a primary care specialty, but the costs of a proceduralist will likely be much higher than an office-based physician.  Were there differences found between specialties defined as primary care?
**The sample sizes are adequate, but I would like to see validity testing in older (65+) populations and underserved populations.
**Except for the previously mentioned issues, validity testing reflects that TCOC is accurately capturing the average expenses per patient.
**Yes. Uses validity testing at both measure score and data element. Uses empirical validity testing.
Some of the questions related to quality do not seem appropriate, as TCC is not necessarily related to quality of provided. This measure can be viewed with quality to get general understanding of value.
**Since it is not clear what the measure is assessing vis a vis resource use, the approach to validity testing is not clear. Overall correlations of the measure with other measures of use or cost demonstrate a weak basis for judging whether the variation in spending is actionable or not, i.e., whether the PMPM cost differences can or should be narrowed.
The risk adjuster is a well-established one, but I would have liked to see more discussion of how much of the variance it explains and the extent to which the rankings change when risk adjustment is introduced.
I would also like to have seen more analysis of variance, breaking allocating the variance in RA PMPM expenses to price differences across groups, and high or low use of specific services.
The high correlation of PMPM and prices suggest that price variations account for a substantial portion of this measure's variance.  How should a payer interpret this? A group?
**Moderate--unclear what measure score over time means.

***2b3. Exclusions Analysis***

Comments:
**Same as resource index --- discuss truncation and exclusions for less than 9 member months.
**Exclusions seem reasonable
**I am bothered that patients under 1 year of age and over 64 years of age are excluded.
**High cost patient are not excluded but are Windsorized (truncated at the threshold value, moving from $100,000 per year to $125,000 per year. This threshold appears to be over 20 times the average cost per person per year. This would still result in potential undue influence by a small number of patients. I would recommend that outliers be excluded rather than "capped". The MSPB measure excludes inter-institution transfers because neither institution has full influence on major portions of the costs incurred. Most cost outliers for TCOC would have the same issues in that much of the cost would be outside the primary care providers control.
**Exclusions are acceptable.
**The Total Cost of Care Population-based PMPM Index measure excludes (truncates) member medical and pharmacy costs that are over $125,000. The AAMC requests an explanation and rationale as to why these medical and pharmacy costs are capped and why $125,000 was selected as the threshold.
The AAMC also requests an explanation as to why non-provider administered drugs (those not covered under Medicare Part B) are not included in the cost calculation for this measure.
**The population exclusions look reasonable.  Risk adjustment is done with a widely adopted metric.
I would like some discussion of the proportion of groups with drug carveouts and the variance in drug spending among those groups for whom the data are available. Would also like to know about other carve outs, particularly mental health services, and the proportion of costs these represent.
**High

***2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures***

Comments:
**Same discussion as resource index measure.
**The risk adjustment presented used the ACG system.  Is that publicly accessible?  I think it is a proprietary tool that has to be licensed to the groups using it.  I am concerned about using an opaque means of risk adjustment in each of these measures. What is used in this measure?  I tried to look this up on the provided website reference, but it just wanted to license the product to me. Did the developers look at the differences in risk adjustment values  between

institutions?
**I would like to know more about the risk adjustment method used. I am concerned that SDS was too readily discarded.
**ACG risk adjustment seems like a reasonable approach for TCOC. It is important to point out that TCOC is developed for a commercially insured population. In that setting, income has little influence. When adding Medicaid patients to the analysis, the reimbursement differences between Medicaid and commercial groups is likely confounded with SDS differences.
**The R2 results further emphasize that ACG score and insurance type are the main drivers of cost and resource use variation and income does not provide any additional predictive power.
**Risk adjustment uses a standard widely adopted measure. The use of census tract level income and other variables is commended. The analysis shows low variance due to SDS variables but this may be due to low variance across the groups included in the analysis of the SDS variables. Would like to see the distribution across groups (not population as a whole) of these measures and better understand the extent of the variance in SDS measures at the group level.
**Moderate--need clarification with the income calculation and its effect on the scores. 1% increase in income results in 0.13 increase in TCC (this looks like large effect on scores)--maybe I'm misreading.

***2b5. Identification of Statistically Significant & Meaningful Differences In Performance***

Comments:
**This measure does not address clinical quality directly
**A total cost of care index is by its very nature a gross, high level measure that does not help identify what factors may have caused the excess cost.
**On pages 32 – 33 in the measure worksheet, performance on the Total Cost Index for most providers falls into a relatively narrow window (between 0.7 and 1.2). In fact, a small number of providers (#s 60-66) appear to be responsible for much of the variation in this measure. Is a quality measure necessarily to address a small number of high spending performers, or are there other more appropriate means to address this issue?
**Recognize the language is standardized, but this is not a quality measure. There appear to be substantial variation in the PMPM costs, with a substantial portion of this due to differences in the prices the groups get.
Would like some committee discussion of the magnitude of the differences across groups.
**Low--not shown. Developer states that statistical significant differences aren't necessary. How do you distinguish provider differences then?

***2b6. Comparability of Performance Scores When More Than One Set of Specifications***

Comments:
N/A

***2b7. Missing Data Analysis and Minimizing Bias***

Comments:
**No
**No
**Does not appear to be a problem.
**high--no missing data to note

## Criterion 3. Feasibility

**Maintenance measures – no change in emphasis – implementation issues may be more prominent**

**3. Feasibility:** This requirement involves demonstrating:

- the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
- the required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.
- the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

**Data Specifications and Elements**

- The measure is constructed using administrative health claims, which are routinely created and do not create undue burden for measure implementers
- All data elements are available in defined fields within electronic sources.

- The measure uses an ACG-Johns Hopkins risk adjustment methodology which is proprietary.

**Data Collection Strategy**

- Data collection strategy can be implemented as it's currently in operational use by HealthPartners

***Questions for the Committee:***

- o *Are the required data elements routinely generated and used during care delivery?*
- o *Are the required data elements available in electronic form EHR or other electronic sources?*
- o *Can the measure be consistently implemented using a proprietary risk adjustment methodology?*

**Preliminary rating for feasibility:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 3: Feasibility**

***3.Feasibility***

Comments:

**The measure uses easily available metrics

**I am concerned that requiring use of a proprietary risk adjustment methodology would reduce widespread use of this measure by increasing implementation costs. Can any existing EMR's calculate this measure? or will additional upgrades or 3rd party software have to be purchased?

**Uses claims data, so generally very feasible to implement.

**Claims based measure. Feasible to implement.

**High--easy to run

## Criterion 4: Usability and Use

**Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences**

**4. Usability and Use**: This requirement involves describing the extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

**Current uses of the measure** [from OPUS]

**Publicly reported?** ☒ **Yes** ☐ **No**

**Current use in an accountability program?** ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

**OR**

**Planned use in an accountability program?** ☐ **Yes** ☐ **No**

**Accountability program details**

- The developer states that there are multiple accountability programs and sub-programs that this measure utilizes including:
  - 3 Public reporting programs
  - 1 Payment program
  - 1 Public Health/Disease Surveillance program
  - 5 Quality Improvement with Benchmarking programs (external benchmarking to organizations)
  - Several Quality Improvement with Benchmarking (internal to the specific organization) programs
- The developer also cited measure page views at the National Quality Measures Clearinghouse (NQMC) from Agency for Healthcare Research and Quality (AHRQ)
  - o Reported the following usage between 3/1/15 – 2/29/16
    - ▪ 5,815 page views for the Total Cost of Care Measure
    - ▪ 1,493 page views for the Total Resource Measure

**Improvement results**

- Large number of those who have adopted the measure and resulted in improvement through greater transparency, which allows users to pinpoint areas for improvement and define strategies to reduce those costs
- One specific example is the Northwest Metro Alliance, which serves more than 300,000 people receiving care at 9 different clinics and one hospital, demonstrated that their medical cost increases were more than 31% lower than the Twin Cities metro average for Commercial patients since they adopted the developer's measure in 2010.

**Unexpected findings (positive or negative) during implementation**

- The developer did not note any unexpected findings during the implementation of the measure

**Potential harms**

- The developer is unaware of negative unintended consequences from other organizations utilizing the measure

**Vetting of the measure by those being measured**

- Since endorsement the measure developer have received some general input regarding implementation of the measure. HealthPartner's organized a public-facing website with resources for external organizations on how to download the necessary tools to run the measure.

**Measure can be deconstructed to facilitate transparency and understanding** ☒ **Yes** ☐ **No**

**Feedback:**

***Questions for the Committee****:*

- o *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- o *Do the benefits of the measure outweigh any potential unintended consequences?*
- o *How has the measure been vetted in real-world settings by those being measured or others?*

**Preliminary rating for usability and use:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 4: Usability and Use**

***4.Usability and Use***

Comments:

**Same discussion on attribution models as resource index measure.

**I would like more information here, again especially considering that this is a maintenance measure. The Developer notes that a number of groups that are collecting data using the measure, but what actual performance data are they seeing? What gaps are being identified? What progress is being made?

**I would like to see this measure vetted in more diverse populations.

**The measure is used in a variety of programs.

- 3 Public reporting programs
- 1 Payment program
- 1 Public Health/Disease Surveillance program
- 5 Quality Improvement with Benchmarking programs (external benchmarking to organizations)
- Several Quality Improvement with Benchmarking (internal to the specific organization) programs

Yes, been vetted in real-world settings by those being measured.

**The developers note that this measure should be used in conjunction with the RCU measure, but the addition this measure offers to understanding of resource use variations is the addition of price. A more useful measure would identify the marginal contribution of price to the PCU measure in explaining variations in resource use, and this is not how the measures are presented. From a user actionability orientation, the reports on these measures provide information by type of service on whether the provider is higher or lower, and this rather than the overall score makes the measure actionable and usable.

**High--of high importance/utility to payers/purchasers (and in turn consumers who they purchase for).

**Criterion 5: Related and Competing Measures**

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

**Related or competing measures**

- The developer did not identify and related or competing measures.

**5.a. Harmonization**: This requirement involves demonstrating that the measure specifications are harmonized with related measures OR the differences in specifications are justified.

N/A

**Endorsement + Designation**

**The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.**

**This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it:** is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

**Eligible for Endorsement + designation**: ☒ **Yes** ☐ **No**

**RATIONALE IF NOT ELIGIBLE**:

## Pre-meeting public and member comments

- Ms. Ellen Gagnon from Network for Regional Healthcare Improvement on 2/21/17:

On behalf of NRHI, we are in support of NQF endorsing this measure. For over three years we have been actively engaged with regions across the country measuring, reporting and using the total resource use population based PMPM index. Recently we published a benchmark report that utilized this measure and compared across 5 regions which has resulted in meaningful conversations within regions about the cause of variation. Seven regions have produced and distributed attributed practice level reports in their communities at least once, some multiple times over the past few years. During 2015, healthcare cost information on over 5 million patients attributed to 20,000 individual physicians were included in practice level reports and used by practices to identify areas of variation and opportunities for intervention to improve care while decreasing costs. The utility of this measure increases as you are able to isolate resource use - which is very powerful and something physicians can control.

The basic foundation for all of these efforts is the HealthPartners NQF endorsed TCOC measure framework. NRHI has been awarded funding from RWJF for a third phase which began on November 1, 2016. During this two-year grant, we will expand the number of regions producing, sharing and using TCOC for both commercial and Medicare populations, maintain and grow our Getting to Affordability Learning Modules and community - a place to connect with others across the country who are measuring and using TCOC, convene a multi-stakeholder summit on using TCOC to advance the Triple Aim and payment reform, and develop and implement sustainability plans to ensure future ability to produce, share and use TCOC.

We support further endorsement of this measure and would be happy to answer any questions.

- Sia Lo on Behalf of Beth Averbeck from HealthPartners Medical Group on 2/22/17:

HealthPartners Medical Group strongly recommends for endorsement both the Total Cost of Care (#1604) and Total Resource Use (#1598) measures. For more than a decade, Total Cost of Care (TCOC) has been the top line measure of affordability for our care group. We drill down from the overall measure of TCOC to price drivers, and Total Resource Use drivers to identify opportunities for improvement. These measures have guided our improvement strategies; allowing us to focus on appropriate use of services and place of service opportunities. This has resulted in improved affordability for our patients. Our full statement of support and usability of these measures was

included in the measure submission.
Nance McClure, Chief Operating Officer and Brian Rank, MD, Executive Medical Director, and Beth Averbeck, MD, Senior Medical Director Primary Care

- Sia Lo on Behalf of Nance McClure from HealthPartners Medical Group on 2/22/17:

HealthPartners Medical Group strongly recommends for endorsement both the Total Cost of Care (#1604) and Total Resource Use (#1598) measures. For more than a decade, Total Cost of Care (TCOC) has been the top line measure of affordability for our care group. We drill down from the overall measure of TCOC to price drivers, and Total Resource Use drivers to identify opportunities for improvement. These measures have guided our improvement strategies; allowing us to focus on appropriate use of services and place of service opportunities. This has resulted in improved affordability for our patients. Our full statement of support and usability of these measures was included in the measure submission.
Nance McClure, Chief Operating Officer and Brian Rank, MD, Executive Medical Director, and Beth Averbeck, MD, Senior Medical Director Primary Care

- Sia Lo on Behalf of Brian Rank, MD from HealthPartners Medical Group on 2/22/17:

HealthPartners Medical Group strongly recommends for endorsement both the Total Cost of Care (#1604) and Total Resource Use (#1598) measures. For more than a decade, Total Cost of Care (TCOC) has been the top line measure of affordability for our care group. We drill down from the overall measure of TCOC to price drivers, and Total Resource Use drivers to identify opportunities for improvement. These measures have guided our improvement strategies; allowing us to focus on appropriate use of services and place of service opportunities. This has resulted in improved affordability for our patients. Our full statement of support and usability of these measures was included in the measure submission.
Nancy McClure, Chief Operating Officer and Brian Rank, MD, Executive Medical Director, and Beth Averbeck, MD, Senior Medical Director Primary Care

- Benson Shih-Han Hsu, MD, MBA, FAAP from Sanford Health on 2/23/17:

Sanford Health supports endorsement of the HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. As an integrated health system in the HealthPartners network, we appreciate the transparency and soundness of the measures, as well as our partnership with HealthPartners as we strive to improve care for our patients. The Sanford Health Plan is also a licensee and user of the measures.

- Steven Mark Connelly, MD from Park Nicollet Health Services on 2/24/17:

Park Nicollet appreciates the opportunity to voice our support for HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. HealthPartners has transparently shared the measurement method and measure results with providers in our community for nearly a decade, and we have used these measures to improve health care affordability for our patients, while maintaining top quality performance. Our full statement of support and comment on the usability and usefulness of these measures was submitted as part of HealthPartners Total Cost of Care and Total Resource Use NQF submission.
Steve Connnelly, MD, President, Park Nicollet Health Services and Kristi Lyon, Vice President, Payer Relations

- Ms. Lori Martin on Behalf of Andrew Dorwart from HealthPartners on 3/1/17:

Stillwater Medical Group and Lakeview Hospital is an integrated, non-profit clinic and hospital system serving the eastern Twin Cities metro area and Western Wisconsin. We use HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598) measures in our system to identify opportunities to improve affordability for our patients. We support maintaining endorsement of the HealthPartners measures.
Andrew Dorwart, MD
Stillwater Medical Group President, Lakeview Hospital System CMO

- Dr. Paul Kasuba from Tufts Health Plan comment on 3/3/17:

Tufts Health Plan supports endorsement of the Health Partners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. These measures have been widely adopted by many stakeholders in the health care community and have advanced the national conversation of health care affordability.

Paul Kasuba, MD SVP/CMO

- Thomas Foels from Independent Health comment on 3/5/17:

Independent Health supports endorsement of HealthPartner's Total Cost of Care (#1604 and #1598) measures. These measures have been adopted by many stakeholders in the health care community and have advanced the national discussion on health care affordability.

- Angelo Sinopoli from Greenville Hospital System comment on 3/5/17:

Greenville Health System fully supports endorsement of the Health Partners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. These measures have been widely adopted by many stakeholders in the healthcare community and have advanced the national conversation around healthcare affordability.
Angelo Sinopoli, MD
VP, Clinical Integration, CMO

- Mr. Akinluwa Demehin, MPH from American Hospital Association comment on 3/6/17:

The American Hospital Association (AHA) recognizes the importance of total cost of care and resource use measures in helping those running health plans better understand and address opportunities to improve the value of the care provided. Therefore, we are exploring a partnership with HealthPartners to pilot use of their measures (#1604 Total Cost of Care and #1598 Total Resource Use), with the goal of using these measures with a subset of our members with health plans to help them better understand their performance. We look forward to working with HealthPartners on designing and implementing this important pilot to enhance value of care for the patients and communities our member organizations serve. Our full letter was included with the HealthPartners submission documents.

- Koryn Y. Rubin from American Medical Association comment on 3/6/17:

Given measure 1598 and 1604 are maintenance measures, the AMA would have expected the developer, HealthPartners to have provided more information on actual performance data and how well the measures performed in the real world across different groups. The developer references all of the groups that started collecting the measure as an indicator that there is progress toward improvement, but uptake of a measure does not mean the same thing as improving performance. We, therefore, have the following concerns:

The measure submission documents state that many groups and institutions are collecting and reporting the measure under the testing and usability section, but we are only provided data from HealthPartner groups in Minnesota and Western Wisconsin. We would like for data from the first submission and anything within the last 4 years to be included and for the data to include mean, std dev, min, max, interquartile range, and scores by decile. It is also not clear to us how HealthPartners standardizes prices.

We also seek clarification on the sample size. The document states it has been tested with a minimum attributed population of 600 members, but it is not clear whether this is with each practice group or by payer or plan. The reliability testing discussion also fails to address the sample size question and the number of physicians or patient that must be attributed to a group for the measure to be considered reliable. This issue was raised as a concern when the measure underwent its last review and once, again, we request more clarity around the level of analysis and how a physician group is defined.

We also find the risk-adjustment strategy utilized for this measure insufficient. The developer utilizes the ACG system which is proprietary and groups must pay to use it. The developer states you can use others but no testing of other risk-adjustment strategies is outlined to compare the results of different tools. It would be helpful to know whether the groups that implemented the measure are all using the ACG system. If not, then it is not quite clear whether the measure produces comparable results across institutions. With the SES analysis, we do not believe the developer provided an adequate conceptual analysis or sufficient information on why they did not test one of the two factors. They first state that they looked at two factors (income and education), cite one or two articles and then they say they could only look at one- income. Therefore, we do not believe what was provided is sufficient to satisfy the SES trial requirements.

We also are concerned with the definition of primary care physician because it includes specialties such as OB/GYN that have higher intensity of services. It would also be helpful to have validity testing that includes comparisons across the different specialties that are defined as primary care physicians by the measure developer and then against all of the groups to see if it can distinguish meaningful differences and not yield inaccurate comparisons by specialty.

- Russ John Kuzel comment on 3/6/17:

SelectHealth supports endorsement of the HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598). These measures have been widely adopted by many stakeholders in the health care community and have advanced the national conversation of health care affordability.

- Sanne Jones Magnan comment on 3/6/17:

Thank you for the opportunity to share my support for the HealthPartners Total Cost of Care (#1604) and Total Resource Use (#1598) measures. With my internal medicine background and my previous leadership roles as the Minnesota Commissioner of Health and President & CEO of the Institute for Clinical Systems Improvement, I know firsthand the importance of the Triple Aim for our communities and our patients. The Total Cost of Care and Total Resource Use measures help leaders, decision-makers, and physicians identify improvement opportunities for affordability and value in our healthcare systems. The measures provide transparent information needed to drive change for better health and experience at a lower cost for our patients and communities.

## Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

**IM1. High Priority**

**IM1.1. Demonstrated High Priority Aspect of Healthcare**

Affects large numbers

High resource use

Patient/societal consequences of poor quality

Severity of illness

**IM1.2. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in IM.1.3.**In 2014, health care spending represented 17 percent of US gross domestic product (GDP); this is the largest percentage of any developed nation in the world.1 A recent survey published by the Commonwealth Fund shows that while the Affordable Care Act has expanded health care coverage, adults in the United States are much more likely to go without needed care because of cost than eleven other westernized countries.2 Consequently, affordability of care continues be highly discussed issue, but in spite of this, prior to 2012, there were few publicly available cost or resource use measures.3,4 Aware of this issue, HealthPartners developed a total cost of care index (TCI) in the late nineties to increase awareness of cost of care and healthcare spending for stakeholders. Total cost reflects a mix of complicated factors including, service utilization, and negotiated prices.3 Non-condition specific cost of care and resource use measures provides valuable information on how to make health care more affordable because health plans and providers can use the data to identify areas where they can lower cost by improving resource use or shifting to less expensive resources (for example, use of a surgery center instead of a hospital where medically appropriate). Evidence supports the idea that improving use of resources and price can lead to lower costs with no loss in quality. Turbyville, et al (2011) found that medical resource use has no relationship with quality of care for diabetes. 5 Fisher, et al (2004) performed a study that showed a similar result for resource use and quality of care in Academic Medical Centers.6 The Medicare Payment Advisory Commission in a report to congress in 2006 also reported that they found no correlation between higher resource use and higher quality of care across six metropolitan statistical areas (MSAs).7

Cost of Care and resource use measures can be used to support a comprehensive measurement system.8 Glass, et al call for reporting of cost and resource use in ACO models as a recommended tool to improve value, they also suggest the use of resources measurement to set targets for payment incentives, by tying payments to quality and resource use improvements.9,10 In addition, overuse of health care services has led to wide variation in health care cost and use across geographies. Studies suggest that Medicare spending would decrease by almost 30 percent if medium and high spending geographies consumed health care services comparable to that of lower spending regions.11 Experts agree that reducing overuse can make care safer and more efficient.12,13 The Total Cost Index, which controls for both cost and illness burden, can be used to identify areas of overuse in health care as well as measure targeted improvement efforts.

**IM1.3. Citations for data demonstrating high priority provided in IM.1.2**

1 The World Bank. Health expenditure, total (% of GDP). http://data.worldbank.org/indicator/SH.XPD.TOTL.ZSend=2014&locations=US&start=1995&view=chart

2 In a New Survey of 11 Countries, US Adults Still Struggle with Access to and Affordability of Health Care. The Commonwealth Fund. November 16, 2016. http://www.commonwealthfund.org/publications/in-the-literature/2016/nov/2016-international-health-policy-survey-of-adults

3 National Committee for Quality Assurance, Insights for Improvement - Measuring Health Care Value: Relative Resource Use, 2010, http://www.ncqa.org/portals/0/hedisqm/RRU/BI%20NCQA_RRU_Publication_FINAL.pdf

4 National Quality Forum. NQF Endorses Resource Use Measures. http://www.qualityforum.org/News_And_Resources/Press_Releases/2012/NQF_Endorses_Resource_Use_Measures.aspx

5 Turbyville, Sally E., Meredith B. Rosenthal, L. Gregory Pawlson, and Sarah Hudson Scholle, Health Plan Resource Use – Bringing Us Closer to Value-Based Decision Making, The American Journal of Managed Care, 2011. Vol. 1, no. 1, p. 68-74. Last accessed http://www.ajmc.com/journals/issue/2011/2011-1-vol17n1/ajmc_2011jan_turbyville_68to74/P-1

6 Fisher, Elliot S., David E. Wennberg, Therese A. Stukel, and Daniel J. Gottlieb, Variations in the Longitudinal Efficiency of Academic Medical Centers, Health Affairs, 2004. doi:10.1377/hlthaff.var.19. http://content.healthaffairs.org/content/early/2004/10/07/hlthaff.var.19.short

7 Medicare Payment Advisory Committee, Report to the Congress: Increasing the Value of Medicare, 2006. http://www.medpac.gov/docs/default-source/reports/Jun06_EntireReport.pdf?sfvrsn=0
8 Fisher, Elliot S.; Shortell, Stephen M. Accountable Care Organizations: Accountable for What, to Whom and How. Journal of American Medical Association. October 20, 2010. http://jama.ama-assn.org/content/304/15/1715.full

9 Glass, David; Stensland, Jeff. Accountable Care Organizations. April 9, 2008. http://medpac.gov/docs/default-source/meeting-materials/april-2008-meeting-transcript.pdf?sfvrsn=0

10.Glass, David; Stensland, Jeff. Accountable Care Organizations. March 12, 2009. http://medpac.gov/docs/default-source/meeting-materials/march-2009-meeting-transcript.pdf?sfvrsn=0

11 Skinner, Jonathan; Fisher, Elliott. The Dartmouth Atlas. Reflections on Geographic Variation in U.S. Health Care. http://www.dartmouthatlas.org/downloads/press/Skinner_Fisher_DA_05_10.pdf

12 National Quality Forum Issue Brief. Waste Not, Want Not: The Right Care for Every Patient. June 2009. www.qualityforum.org/Publications/2009/07/Waste_Not_Want_Not_Issue_Brief.aspx

13 National Priorities and Goals. National Priorities Partnership convened by the National Quality Forum. November 2008. https://www.qualityforum.org/Setting_Priorities/NPP/National_Priorities_Partnership_Goals.aspx

IM2. **Opportunity for Improvement**

IM2.1. **Briefly explain the rationale for this measure (e.g., the benefits or improvements in performance envisioned by use of this measure)**

By measuring population based total cost of care, health plans and providers can improve the affordability of health care without sacrificing quality. HealthPartners' TCI gives provider groups valuable information on the cost of care and, when viewed in conjunction with resource use and quality metrics, information on the efficiency of care. The HealthPartners TCI measure is a population-based, patient-centered, total cost of care measure that crosses all categories of health services. This is in contrast to the many, episodic based measures available in the market today. Both population based and episodic based measures are important and complementary but a key benefit of population based measures is helping to better understand potential overuse & underuse (e.g., although efficient at spine surgery, may be performing too many).

IM2.2. **Provide performance scores on the measure as specified** (current and over time) **at the specified level of analysis.** (This is required for endorsement maintenance. Include mean, stddev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include). **This information also will be used to address the subcriterion on improvement (U.2.1.) under Usability and Use.**

The Dartmouth Atlas has been an eye-opening look at the variation in health care spending and resource use across regions for the

Medicare population. The measurement of cost of care and resource use is as widely varied in the commercial population across geographies.1 While HealthPartners has applied the measure on the commercial population, the measure could easily be applied to other populations.

A study of the Minnesota market further highlighted the significant variation in cost and efficiency ranging from $2,400 to $4,700 PMPY. Additional findings found no relation to quality or type of practice (large, small, integrated, etc).2 These findings are further confirmed based on HealthPartners own experience and analyses. Existing total cost and resource use measures are largely condition or episode specific measures. Prior to 2012, there was not an existing total population cost of care measure in the market that crossed all care services.3 A Total Cost of Care measure was implemented by the Integrated Healthcare Association in California. 4. Based on 2015 dates of service, the multi-stakeholder community collaborative, Minnesota Community Measurement (MNCM) measured the Total Cost of Care of 257 provider groups, representing 1.5 million patients receiving care. The data were source from the four major commercial payer in Minnesota. The 2015 risk-adjusted total cost of care per member per month on average was $474, with a range of $365 to $916. Eighty percent of provider groups were between $394 and $555 per member per month.5

IM2.3. **If no or limited performance data on the measure as specified is reported in IM.2.2., then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

1.Dartmouth Atlas. http://www.dartmouthatlas.org/
2.Kralewski, John E, Dowd, Bryan E, Xu, Yi (Wendy). Differences in the Cost of Health Care Provided by Group Practices in Minnesota. February 2011. Minnesota Medicine. http://www.minnesotamedicine.com/tabid/3678/Default.aspx
3.Berwick, Donald M., Nolan, Thomas W., Whittington, John, The Triple Aim: Care, Health and Cost. Health Affairs, May/June 2008.
doi: 10.1377/hlthaff.27.3.759. http://content.healthaffairs.org/content/27/3/759.full?sid=f3d381e8-76ef-415f-9080-de97c1273fa6
4.Integrated Healthcare Association (IHA) Total Cost of Care. Measuring and Using Total Cost of Care Data in California. Fact Sheet. http://www.iha.org/sites/default/files/resources/fact-sheet-total-cost-of-care-2016.pdf
5. Minnesota Community Measurement. 2016 Cost and Utilization Report: Average Cost per procedure, Total Cost of Care, Relative Resource Use, Utilization. http://mncm.org/wp-content/uploads/2016/12/16CostUtilityReport.pdf

IM2.4. **Provide disparities data from the measure as specified** (current and over time) **by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** (This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) **This information also will be used to address the subcriterion on improvement (U.2.1.) under Usability and Use.**

As previously described in this application, the measure is being submitted for a commercially insured population. Therefore performance by insurance status is not applicable because the population is all commercially insured. The clinical risk adjustment process described in 2b4.3 describes how age and gender are accounted for in the methodology and no additional measure performance was tested because this is not how they are being used. That said, in looking at single specialty obstetric and pediatric groups, we see a uniformly distributed result across our network performance and these groups are not clustered, which demonstrates results are not biased against age or gender. Additionally, this demonstrates the clinical risk adjustment is working effectively. The measure is used as a population-based method primarily for payment, benefit design, transparency and improvement.

After applying clinical risk adjustment, socioeconomic testing was conducted that considered income and education status as potential factors beyond those already adjusted for.

Model Results

1% Income Increase:
Total Reimbursement $(0.13)
Resource Use $0.16
Price $(0.28)

1% ACG Increase:
Total Reimbursement $4.22
Resource Use $4.34
Price $(0.07)

Commercial vs. Medicaid Membership:
Total Reimbursement $133.28
Resource Use $(75.24)
Price $205.36

Resource Use Endorsed Measure R2 = 0.5788
Resource Use Endorsed Measure + Income R2 = 0.5792

Using Census tract data, a 1% increase in income resulted in a $0.13 decrease in total reimbursement, a $0.16 increase in resource use, and $0.28 decrease in price. The results highlight how significantly more the ACG score (clinical risk adjustment) and insurance product impact both the cost and resource use measures. For frame of reference, on average for the Midwest market, the total spend for a member per month (PMPM) is $400. The results of the evaluation show that a 1% increase in risk score accounts for a $4.22 or roughly 1% increase in PMPM.
Product also contributed significantly with there being a $133 dollar difference in cost between commercial and Medicaid. The variation in resource use was much less, however, still significant with Medicaid covered members utilizing $75 more dollars of resources. The fact that Medicaid's cost per service is approximately half that of commercial rates drives the differences between the TCOC and Resource Use results. The R2 results further emphasize that ACG score and insurance type are the main drivers of cost and resource use variation and income does not provide any additional predictive power.
Methodology and testing results can be found here:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

IM2.5. **If no or limited data on disparities from the measure as specified is reported in IM.2.4., then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Not applicable

IM3. **Measure Intent**

IM3.1. **Describe intent of the measure and its components/ Rationale (including any citations) for analyzing variation in resource use in this way.**
Key considerations when constructing the measure:
• The purpose of population-based measurement is to better understand overuse, underuse, and person-centered management and accountability
• Population based-measurement nicely complements condition and episode-base measures, combined they depict a complete picture of total cost of care.
• Risk adjustment is a critical component to the measure to allow for fair comparisons
• Use this measure as part of a Triple-aim approach where the Total Cost of Care measure complements resource use, quality and patient experience.
• The Total Cost Index measure when used with a Resource Use Index measure helps to better understand cost and resource use opportunities.

## Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

**Specifications** The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Non-Condition Specific** *(check all the areas that apply):*
Care Coordination
Safety : Overuse

De.7. **Care Setting** *(Select all the settings for which the measure is specified and tested):*
Ambulatory Surgery Center
Behavioral Health : Inpatient
Behavioral Health : Outpatient
Birthing Center
Clinician Office/Clinic
Dialysis Facility
Emergency Department
Emergency Medical Services/Ambulance
Home Health
Hospice
Hospital
Hospital : Acute Care Facility
Hospital : Critical Care
Imaging Facility
Inpatient Rehabilitation Facility
Laboratory
Long Term Acute Care
Nursing Home / SNF
Other:All care settings
Outpatient Rehabilitation
Pharmacy
Urgent Care - Ambulatory

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
For purposes of resubmission please use the following link to view materials including updated measure specifications: www.healthpartners.com/tcoc-documents  For reference, currently endorsed measure materials reside here: www.healthpartners.com/tcoc

S.2. **Type of resource use measure** *(Select the most relevant)*
Per capita (population- or patient-based)

S.3. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED):*
Clinician : Group/Practice, Population : Community, County or City

S.4. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*

S.5. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
*If other, please describe in S.5.1.*
Claims (Only)

S.5.1. **Data Source or Collection Instrument** *(Identify the specific data source or data collection instrument, e.g. name of database, clinical registry, collection instrument, etc.)*
Use administrative claims data base

Risk Adjustment Tool, Johns Hopkins ACG System

S.5.2. **Data Source or Collection Instrument Reference** *(available at measure-specific Web page URL identified in S.1 OR in the file attached here) (Save file as: S_5_2_DataSourceReference)*

S.6. **Data Dictionary or Code Table** *(Please provide a web page URL or attachment if exceeds 2 pages. NQF strongly prefers URLs. Attach documents only if they are not available on a web page.)*

***Data Dictionary:***

URL:

Please supply the username and password:

Attachment:

***Code Table:***

URL:

Please supply the username and password:

Attachment:

**Construction Logic**

S.7.1. **Brief Description of Construction Logic**

If applicable, summarize the general approach or methodology to the measure construction. This is most relevant to measures that are part of or rely on the execution of a measure system or applies to multiple measures.

The measure examines total cost of care of a commercial population for a given measurement year (e.g. January 1 and December 31), for all members eligible for the measure.

S.7.2. **Construction Logic** *(Detail logic steps used to cluster, group or assign claims beyond those associated with the measure's clinical logic.)*

- All claims included in the measure have a date of service in the measurement year (e.g. between January 1 and December 31)
- Members have a minimum 9 months enrollment in the measurement year
- Commercial population only
- Attribution
- Risk Adjustment

S.7.2a. **CONSTRUCTION LOGIC ATTACHMENT or URL:** If needed, attach supplemental documentation (Save file as: S_7_2_Construction_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.

URL: https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_187908.pdf

Please supply the username and password:

Attachment:

S.7.3. **Concurrency of clinical events, measure redundancy or overlap, disease interactions** *(Detail the method used for identifying concurrent clinical events, how to manage them, and provide the rationale for this methodology.)*

We do not provide specifications for concurrency of clinical events.

Not applicable. This is a population-based measure that applies to all care settings and conditions.

S.7.4. **Complementary services** *(Detail how complementary services have been linked to the measure and provide rationale for this methodology.)*

We do not provide specifications for linking complementary services.

Not applicable. This is a population-based measure that applies to all care settings and conditions.

**S.7.5. Clinical hierarchies** *(Detail the hierarchy of codes or condition groups used and provide rationale for this methodology.)*

We do not provide specifications for clinical hierarchies.

**S.7.6. Missing Data** *(Detail steps associated with missing data and provide rationale for this methodology (e.g., any statistical techniques to impute missing data)*

We do not provide measure specifications or guidelines for missing data :

In the instances where members have pharmacy benefit carve-outs the following methodology is applied.

The Total Cost of Care measure accounts for members that have their pharmacy benefit carved out by using the members that have pharmacy coverage as a proxy. This technique allows for members without pharmacy coverage to be included in the medical portion of the total cost of care with their pharmacy costs reflecting the provider's risk adjusted pharmacy costs from those covered. The measures separate the total spend into medical and pharmacy and only includes the members with pharmacy coverage into the PMPM calculation for pharmacy. The total PMPM for a provider group is then calculated by adding the medical PMPM to the pharmacy PMPM: Total PMPM = (Medical Costs / Medical MMs) + (Pharmacy Costs / Pharmacy MMs). MM = member months.

HealthPartners' data includes all medical and mental health care. It also includes the majority of pharmacy claims with the exception of some carveouts. The methodology described above was used for testing. If users have additional carve-outs (e.g., mental health) the lowest common denominator principle (i.e. for any given user if their data includes a carve-out for one their method must apply a carve-out for all) needs to be applied to ensure providers are evaluated fairly. This would require all services that are carved out of one segment of input data to be carved out of the measure for all segments of input data and all input components of the measure (e.g. PMPMs, attribution, and risk adjustment).

**S.7.7. Resource Use Service Categories (Units) (Select all categories that apply)**

Inpatient services: Inpatient facility services

Inpatient services: Evaluation and management

Inpatient services: Procedures and surgeries

Inpatient services: Imaging and diagnostic

Inpatient services: Lab services

Inpatient services: Admissions/discharges

Inpatient services: Labor (hours, FTE, etc.)

Other inpatient services

Ambulatory services: Outpatient facility services

Ambulatory services: Emergency Department

Ambulatory services: Pharmacy

Ambulatory services: Evaluation and management

Ambulatory services: Procedures and surgeries

Ambulatory services: Imaging and diagnostic

Ambulatory services: Lab services

Ambulatory services: Labor (hours, FTE, etc.)

Other ambulatory services

Durable Medical Equipment (DME)

Other services not listed

All care is included

All care is included

All care is included

S.7.8. **Identification of Resource Use Service Categories (Units)**
*(For each of the resource use service categories selected above, provide the rationale for their selection and detail the method or algorithms to identify resource units, including codes, logic and definitions.)*
The Total Cost of Care considers 100% of health care services in the Total Cost Index and is calculated on a risk-adjusted paid per member per month basis as well as benchmarked to a peer group. The paid amount (i.e., allowed) is inclusive of both plan and member liability.

S.7.8a. **If needed, provide supplemental resource use service category specifications in either URL (preferred) or as an attachment (Save file as S.7.8a_RU_Service_Categories):**
URL:

Please supply the username and password:

Attachment:

**Clinical Logic**

S.8.1. **Brief Description of Clinical Logic** (Briefly describe your clinical logic approach including clinical topic area, whether or not your account for comorbid and interactions, clinical hierarchies, clinical severity levels and concurrency of clinical events.)
Not applicable. This is a population-based measure that applies to all care settings and conditions.

S.8.2. **Clinical Logic** *(Detail any clustering and the assignment of codes, including the grouping methodology, the assignment algorithm, and relevant codes for these methodologies.)*
Not applicable. This is a population-based measure that applies to all care settings and conditions.

S.8.3. **Evidence to Support Clinical Logic Described in S.8.2** *Describe the rationale, citing evidence to support the grouping of clinical conditions in the measurement population(s) and the intent of the measure (as described in IM3)*
Not applicable. This is a population-based measure that applies to all care settings and conditions.

S.8.3a. **CLINICAL LOGIC ATTACHMENT or URL: If needed, attach <u>supplemental</u> documentation (Save file as: S_8_3a_Clinical_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.**
URL:

Please supply the username and password:

Attachment:

S.8.4. **Measure Trigger and End mechanisms** *(Detail the measure's trigger and end mechanisms and provide rationale for this methodology)*
All claims dates of service in the measurement year (e.g. January 1 – December 31).

S.8.5. **Clinical severity levels** *(Detail the method used for assigning severity level and provide rationale for this methodology)*

We do not provide specifications for clinical severity levels.

This is accounted for in application of risk adjustment, Johns Hopkins, ACG System

S.8.6. **Comorbid and interactions** *(Detail the treatment of co-morbidities and disease interactions and provide rationale for this methodology.)*
We do not provide specifications for co-morbidies and disease interactions.

This is accounted for in application of risk adjustment, Johns Hopkins, ACG System

**Adjustments for Comparability**

S.9.1. **Inclusion and Exclusion Criteria** *Detail initial inclusion/exclusion criteria and data preparation steps (related to clinical exclusions, claim-line or other data quality, data validation, e.g. truncation or removal of low or high dollar claim, exclusion of ESRD patients)*

We do not provide measure specifications or guidelines for data inclusion criteria :
The HealthPartners' Total Cost of Care measure is a full population-based measure, with members under age 1, members 65+ and members with less than 9 months of enrollment excluded to ensure an accurate risk assessment is made on the population.

- Members over age 64
- Members under age 1
- Member enrollment less than nine months during the one year measurement time window
- Dollars per member up to $125,000 are included; dollars per member above $125,000 are excluded (truncated)

- Administrative claims covering all categories of health care services: professional, facility inpatient and outpatient, pharmacy, lab, radiology and any other ancillary healthcare services
- Johns Hopkins ACG System for risk adjustment
- Membership eligibility, identifier and number of months during the measurement period the member was eligible (member months)

The following should be reviewed prior to beginning implementation of the Total Cost of Care measure to ensure data comparability:

- Consistent population of primary and secondary claims diagnosis. Population prevalence to ensure reasonable/completeness of disease; primary and secondary diagnosis are consistently populated (e.g., diagnosis 1 - 4)
- Data elements are populated within reasonable tolerances and thresholds (e.g., expected CPT ranges, expected allowed amount ranges, expected units ranges)
- All service categories are available and appropriately represented (e.g., inpatient, pharmacy, outpatient and professional)
- Peer group/case-mix need to be comparable
- Risk adjustment weight and application must be in sync (e.g. truncation threshold values)

It is recommended that further reliability and validity testing be conducted if the user varies from the "Technical Guidelines" provided. Examples include:

- The user implements the measure with less than 600 members attributed to a provider
- The user applies a different unit of evaluation, such as an employer group, condition or community rather than a provider
- The user employs an alternative attribution algorithm or risk adjustment tool

Paid medical and pharmacy administrative claims for the measurement year (e.g. between January 1 and December 31), allowing for three months of run out for claims lag.

In the instances where members have pharmacy benefit carve-outs the following methodology is applied.
The Total Cost of Care measure accounts for members that have their pharmacy benefit carved out by using the members that have pharmacy coverage as a proxy. This technique allows for members without pharmacy coverage to be included in the medical portion of the total cost of care with their pharmacy costs reflecting the provider's risk adjusted pharmacy costs from those covered. The measures separate the total spend into medical and pharmacy and only includes the members with pharmacy coverage into the PMPM calculation for pharmacy. The total PMPM for a provider group is then calculated by adding the medical PMPM to the pharmacy PMPM: Total PMPM = (Medical Costs / Medical MMs) + (Pharmacy Costs / Pharmacy MMs). MM = member months.

HealthPartners' data includes all medical and mental health care. It also includes the majority of pharmacy claims with the exception of some carveouts. The methodology described above was used for testing. If users have additional carve-outs (e.g., mental health) the lowest common denominator principle (i.e. for any given user if their data includes a carve-out for one their method must apply a carve-out for all) needs to be applied to ensure providers are evaluated fairly. This would require all services that are carved out of one segment of input data to be carved out of the measure for all segments of input data and all input components of the measure (e.g. PMPMs, attribution, and risk adjustment).

S.9.2. **Risk Adjustment Type** (Select type)
Statistical risk model
If other:

S.9.3. **Statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables.)*
For the Total Cost of Care measurement, risk adjustment is performed using Adjusted Clinical Groups (ACG) developed by Johns

Hopkins University. The Johns Hopkins ACG® System has the distinction of being developed, tested and supported by a world renowned
academic and medical research institution, The Johns Hopkins University. The academic home of the ACG System allows for an unparalleled openness to the method. Each component of the system is exposed to the user which allows the system to be easily adapted to unique local circumstances and applications. The ACG methodology is subject to continuous critical review and
testing by a team of distinguished health services researchers led by Dr. Jonathan Weiner. This transparency and academic credibility is critical when trying to disseminate risk information to providers and purchasers of healthcare.
Attributed members are assigned a risk score based on diagnoses on claims from the performance measurement period, as well as member age and gender. The Society of Actuaries Accuracy of Claims-Based Risk Scoring Models (2016) findings suggest other comparable risk groupers are available and would need to be tested for reliability and validity of that risk grouper.
https://www.soa.org/Files/Research/research-2016-accuracy-claims-based-risk-scoring-models.pdf

For the purpose of this application, this measure has been tested using the Johns Hopkins University developed Adjusted Clinical Groups (ACG System).

http://acg.jhsph.org/

Technical Paper: https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/dev_057425.pdf

Risk Adjustment Specifications
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/dev_057913.pdf

ACG Technical Guide
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf

See Measure Testing Attachment for more information on the statistical risk model method and variables

S.9.4. **Detailed Risk Model Specifications** *available at measure-specific Web page URL identified in S.1 OR in attached data dictionary/code list Excel or csv file.*
Available at measure-specific web page URL identified in S.1

S.9.5. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets)*
Measures are adjusted for clinical risk and limited to the commercial population.

S.9.6. **Costing method**
Detail the costing method including the source of cost information, steps to capture, apply or estimate cost information, and provide rationale for this methodology.
Actual prices paid
The Total Cost of Care considers 100% of health care services in the Total Cost Index and is calculated on a risk-adjusted paid per member per month basis as well benchmarked to a peer group. The paid amount (i.e., allowed) is inclusive of both plan and member liability.

S.10. **Type of score***(Select the most relevant):*
Ratio
Other (specify):
If other: https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/dev_057910.pdf see page 9
Attachment:

S.11. **Interpretation of Score** *(Classifies interpretation of a ratio score(s) according to whether higher or lower resource use amounts is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score, etc.)*
A provider Total Cost Index (TCI) of 1.10 equates to 10% higher paid risk adjusted PMPM. Similarly, a provider TCI score of 0.90 equates to 10% less paid risk adjusted PMPM.
A score of 1.0 is equivalent to the peer group average.

S.12. **Detail Score Estimation** *(Detail steps to estimate measure score.)*
There is no estimation in the Total Cost of Care Measure.  The actual result is calculated as follows:
Total Cost Index (TCI):
Numerator: Total PMPM = (Total Medical Cost / Medical Member Months) + (Total Pharmacy Cost / Pharmacy Member Months)

Denominator:  Average Risk Score - the medical claims data is submitted through the Johns Hopkins ACG Risk Grouper which generates a relative risk score for each member. That risk score is then multiplied by the number of months a member has been enrolled creating a risk weight. The risk weights are then summed to the desired level of measurement (e.g., provider group) and divided by the total sum of the desired level's member months creating a member month weighted Average Risk Score.

ACG Adjusted PMPM = Total PMPM / ACG Risk Score
TCI = Provider ACG Adjusted PMPM / Peer Group ACG Adjusted PMPM

**Reporting Guidelines**
This section is optional and will be available for users of the measure as guidance for implementation and reporting.

S.13.1. **Describe discriminating results approach**

Detail methods for discriminating differences (reporting with descriptive statistics--e.g., distribution, confidence intervals).
This is a full population-based measure, therefore, confidence intervals are not applicable. The results can be analyzed by percentile, percent from mean, standard deviation and clustering methods, this is dependent upon the business application of the
measure.

A provider Total Cost Index (TCI) score of 1.10 equates to 10% more cost than the peer group average. Similarly, a provider TCI score of 0.90 equates to 10% less cost than the peer group average. A score of 1.00 is equivalent to the peer group average.

S.13.2. **Detail attribution approach**

Detail the attribution rules used for attributing resources/costs to providers (e.g., a proportion of total measure cost or frequency of visits during the measure's measurement period) and provide rationale for this methodology.
To determine which members to include in the Total Cost of Care measure, there are several options available depending upon your business purpose and unit of measure. The unit of measure could be an entire health plan, provider group, employer group and/or geographic in nature.

Measure was tested using commonly used Attribution Algorithm in an open access market (plurality model, using most recent visit as a tie breaker):
- Include twelve months based on first date of service for the measurement year (e.g. January 1 – December 31) of professional claims experience, with three months of paid claims run out to allow for claims lag.
- Exclude all services that are not office based
- Exclude convenience care clinic visits and hospice services
- Exclude a providers that are not a physician, physician assistant or nurse practitioner
- Assign each service line a specialty based on the servicing physician's practicing specialty or credential specialty if practicing specialty is not available.
- Include only the following specialties:

- Family Medicine, Internal Medicine, Pediatrics, Geriatrics, OB/GYN

Technical specifications:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/dev_057425.pdf

HealthPartners has studied various attribution methods, our findings are located here: HealthPartners Attribution Technical Paper
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_031064.pdf

S.13.3. **Identify and define peer group**
Identify the peer group and detail how peer group is identified and provide rationale for this methodology.
The peer group can be applied by market, region or national with the following criteria:

• Provider Specialties include: Internal Medicine, Family Medicine, Pediatrics, Geriatrics and OB/GYN
• Provider Types include: Physician, Physician Assistant, Nurse Practitioner

S.13.4. **Sample size**

Detail the sample size requirements for reporting measure results.

This measure has been tested for a minimum attributed member population of 600 members, this number is aligned with over 80+ community-based quality and patient experience measures in the market tested. We recommend further reliability and validity testing if a threshold less than 600 attributed members is used.

S.13.5. **Define benchmarking and comparative estimates**

Detail steps to produce benchmarking and comparative estimates and provide rationale for this methodology.

The Total Cost of Care measure is relative to a benchmark or peer group of the user's choice. This can be a group of members or providers, geographic regions or any grouping of member data. The idea is that the Total Cost of Care measure will return a value that will be relative to the peer group average (e.g., 1.10 = 10% higher than the peer group average).

The peer group average is set as the benchmark and a provider's Total Cost of Care ACG Adjusted PMPM indexed against the peer group average. The Peer Group average is calculated in the same manner as an individual provider:
Total Cost (TCI):
Numerator: Peer Group Total PMPM = (Peer Group Total Medical Cost / Peer Group Medical Member Months) + (Peer Group Total Pharmacy Cost / Peer Group Pharmacy Member Months)

Denominator: Peer Group ACG Risk Score Peer Group ACG Adjusted PMPM = Peer Group Total PMPM / Peer Group ACG Risk Score

Total Cost Index: TCI = Provider ACG Adjusted PMPM / Peer Group ACG Adjusted PMPM

**Validity – See attached Measure Testing Submission Form**

**SA.1. Attach measure testing form**

NQF_testing_attachment_Total_Cost_of_Care_1604_021517-636227630530340045.docx

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

**Measure Number** (*if previously endorsed*)**:** 1604

**Measure Title**: Total Cost of Care Population-based PMPM Index

**Date of Submission**: 12/1/2016

**Type of Measure:**

| | |
|---|---|
| ☐ Outcome (*including PRO-PM*) | ☐ Composite – ***STOP – use composite testing form*** |
| ☐ Intermediate Clinical Outcome | X Cost/resource |
| ☐ Process | ☐ Efficiency |
| ☐ Structure | |

**Instructions**

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.

- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

**Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.**

**2a2. Reliability testing** [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

**2b2. Validity testing** [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

**2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
**AND**
If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

**2b4. For outcome measures and other measures when indicated** (e.g., resource use):
- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14],[15] and has demonstrated adequate discrimination and calibration

**OR**
- rationale/data support no risk adjustment/ stratification.

**2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;
**OR**
there is evidence of overall less-than-optimal performance.

**2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results**.

**2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

**Notes**
**10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
**11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data

elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

**12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

**13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

**14.** Risk factors that influence outcomes should not be specified as exclusions

**15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

HealthPartners has developed a Total Cost of Care (TCOC) measure and a Total Resource Use measure. The two measures use the same measurement criteria except for the costing method. While the measures can be used independently, when used together they provide a comprehensive evaluation of cost and further identify opportunities to improve affordability. TCOC measure is a combination of resource use and price and measures the cost effectiveness of managing a population. Total Resource Use measure removes price and measures the frequency and intensity of services.

Because Resource Use is a component of Total Cost of Care, the two measures are complementary to each other, therefore the two measures are tested and evaluated together for reliability and validity, also increasing efficiency of testing by the measure developer. References to both measures are included in the links to technical papers and table of results found throughout the attachment.

Note: Information from prior submission in 2012 is included in *gray italic font* within the body of the form. Methodology used for testing remains the same as prior submission. Results from prior testing are included as a packaged PDF of technical papers within Appendix A. The packaged reports provide a complete analytical pathway with context and reasoning to conclude the measure is reliable and valid.

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

| **Measure Specified to Use Data From:**<br>**(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
|---|---|
| ☐ abstracted from paper record | ☐ abstracted from paper record |
| x☐ administrative claims | x☐ administrative claims |
| ☐ clinical database/registry | ☐ clinical database/registry |
| ☐ abstracted from electronic health record | ☐ abstracted from electronic health record |
| ☐ eMeasure (HQMF) implemented in EHRs | ☐ eMeasure (HQMF) implemented in EHRs |
| ☐ other: Click here to describe | ☐ other: Click here to describe |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Commercial administrative claims
Medicaid administrative claims were used in addition to commercial claims for purposes of socio-economic status (SES) testing.

**1.3. What are the dates of the data used in testing**? Click here to enter date range

2014, 2015 dates of service for validity testing
2015 dates of service for reliability testing
2015 dates of service for SES testing

*Prior submission: 2007, 2008, 2009 dates of service*

**1.4. What levels of analysis were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

| **Measure Specified to Measure Performance of:**<br>(***must be consistent with levels entered in item S.26***) | **Measure Tested at Level of:** |
|---|---|
| □ individual clinician | □ individual clinician |
| x□ group/practice | x□ group/practice |
| □ hospital/facility/agency | □ hospital/facility/agency |
| x□ health plan | □ health plan |
| □ other: Click here to describe | □ other: Click here to describe |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

HealthPartners primary care network (Minnesota and western Wisconsin) consists of 66 individual provider groups that have 850 clinic sites. Provider group size vary from 600 to a few large systems with 40,000+ members.

*Prior submission: HealthPartners' primary care Twin Cities metro area providers as per the specifications of the measure for the calendar years of 2007, 2008 and 2009. HealthPartners primary care metro network consists of 19 individual providers that have 223 (2007) 232 (2008) and 229 (2009) clinic sites.*

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

This is a population-based measure that applies to all care settings and conditions using HealthPartners health plan's full book of business. The total membership of the primary care attributed network is over 530,000 members in 2015.

*Prior submission: The total membership of the primary care attributed metro network membership grew slightly over the three year period: 268,912 (2007), 272,491 (2008) 303,638 (2009).*

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

Reliability and Validity testing use the same population and underlying data. The SES testing also includes the Medicaid population.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

The Total Cost of Care measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). The measure is also limited to commercial only. Socioeconomic testing was conducted that considered income and education status as potential factors beyond those already adjusted for.

___________________________________

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

x☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

*Prior submission: Please see Appendix A (page 2) for reliability testing results from prior submission. The method of testing (bootstrapping and 90% random sample) used for current resubmission is the same methodology used in prior submission.*

**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Overview of Analysis

Total Cost of Care Total Cost Index (TCI) is a measure of a provider's effectiveness of managing their primary care attributed population across the care continuum. The TCI measure was applied to HealthPartners primary care providers as per the measure specifications and results were calculated for 2015.

The reliability testing demonstrates the repeatability of producing the same results a high proportion of the time. To measure the reliability of the TCI measure the actual results were compared to the results calculated by two sampling methods, bootstrapping and a 90% random sample.

These methods were chosen as they represent the measure intent, which is that the TCI measure represents providers' average total cost of care across their population. Since the measure is aggregated to the provider group level, evaluation of member level variability is not necessary.

In the bootstrapping method members that were attributed to a provider group were randomly selected with replacement. This method artificially creates variation around a provider group's total cost of care as each randomly selected iteration (sample populations) does not truly represent the provider's case mix of patients. What this method does however is give an indication as to the repeatability of the measure by comparing how closely the actual total cost measure is to the bootstrapped averages.

In the 90% random sample method, the members that were attributed to a provider group were randomly sampled at the 90% membership level without replacement. This technique was employed to create variation within a provider group by leveraging their own population and controlling for the patient case mix variation that is introduced when random sampling is employed.

Methodology

To perform the bootstrap, the SAS PROC SURVEYSELECT procedure with the Unrestricted Random Sample option for full replacement was utilized to create a series of random samples for each provider group being measured. Full replacement means that one observation is drawn at random, recorded, and then placed back into the data pool so that it can be drawn again if randomly selected. The numbers of records sampled are drawn such that the samples created are the same size as the original number of attributed members for the provider group. In this way, it is theoretically possible (although virtually improbable) to produce a sample of size n that could consist of the same record drawn n times in a row. This sample process was performed 500 times for each provider group being analyzed, to produce 500 sets of risk-adjusted Total Cost of Care results for each provider included in the analysis.

Once the 500 samples were created for each provider group, the total costs of care of each sample for each provider group were compared to the network average to produce risk adjusted indices. The mean Total Cost Index (TCI) from these 500 iterations was computed and compared to the Actual TCI index for each provider group.

In the second method, 90% of attributed provider group members were randomly selected, without replacement. A 90% sample was used despite having the full health plan provider population, as a concession to provider claims that errors in administrative data may not allow for a perfect 100% representation of their population. The sampling process was performed using the SAS PROC SURVEYSELECT procedure with the Simple Random Sample (SRS) option. This method allows for each attributed member to be selected only one time until 90% of the total provider population has been reached. The 90% sampling process was repeated 500 times for each provider group analyzed. Attributed members' total costs were aggregated in each sample to produce 500 Total Cost Index results for each provider group. The mean of the sampled Total Cost index was calculated for each provider group and compared to the Actual TCI index for each provider group.

The bootstrap results should indicate that the within provider TCI variation is significantly less than the between provider variation.

Reliability Paper includes the same method of testing described above:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188105.pdf

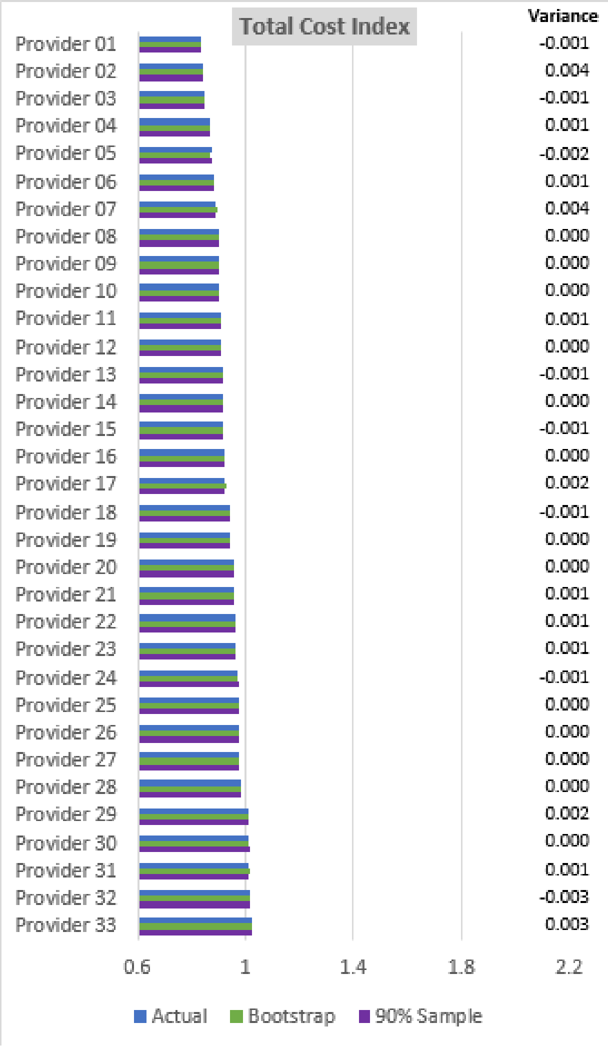
**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (*e.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

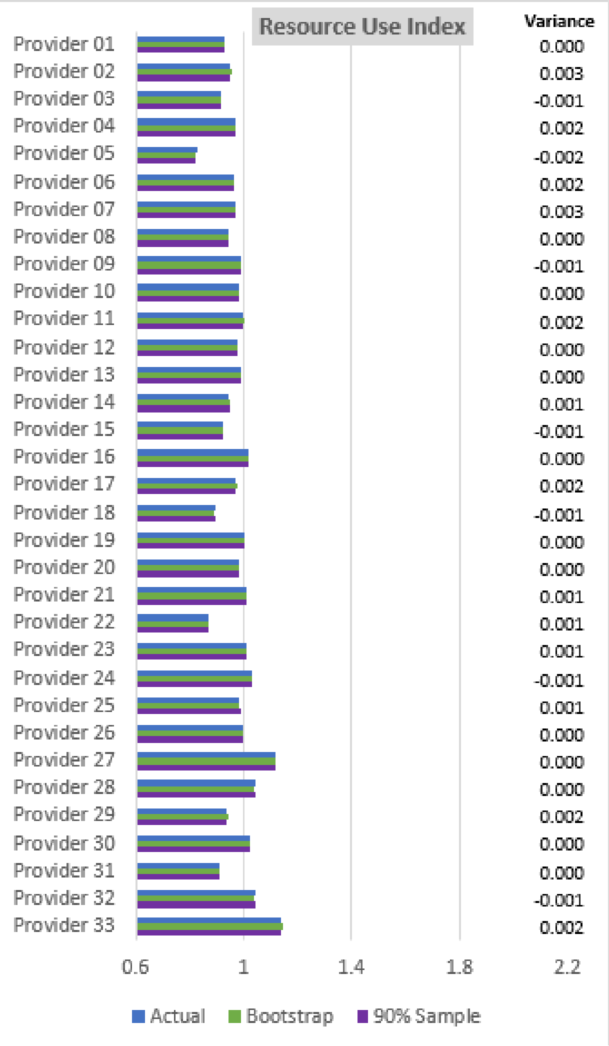
- The differences between the Actual TCI results and both the bootstrap and 90% sample results are very small ranging from -0.0059 to 0.0075 in the bootstrap to -0.0022 to 0.0012 in the 90% sample.

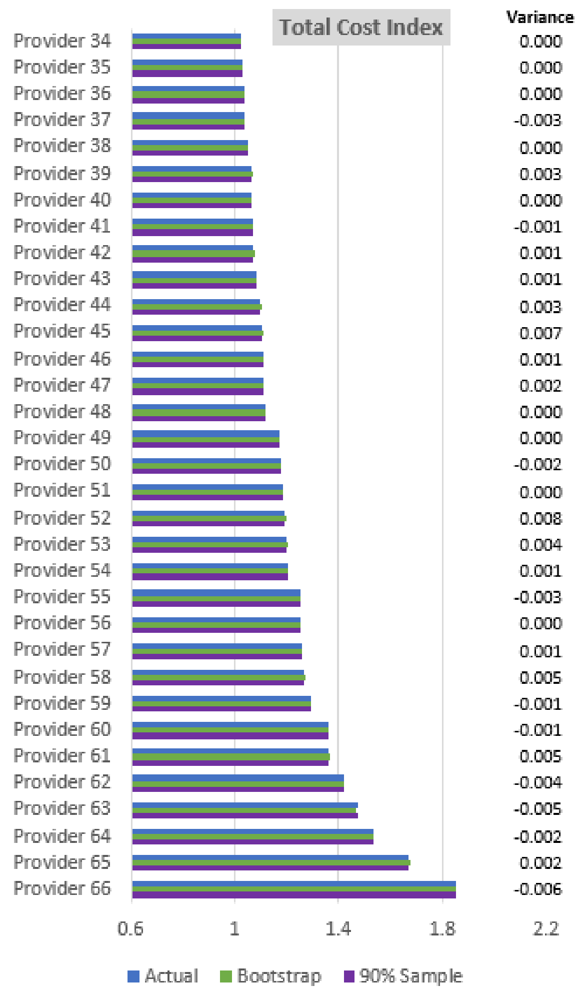
The mean Total Cost of Care results from the bootstrap and 90% samples compared to the actual TCI results for each provider group are displayed on the following charts. The variance between the actual TCI to the bootstrap results is shown on the far right of each chart. The charts are sorted in ascending order by TCI as referenced in the Reliability Paper.

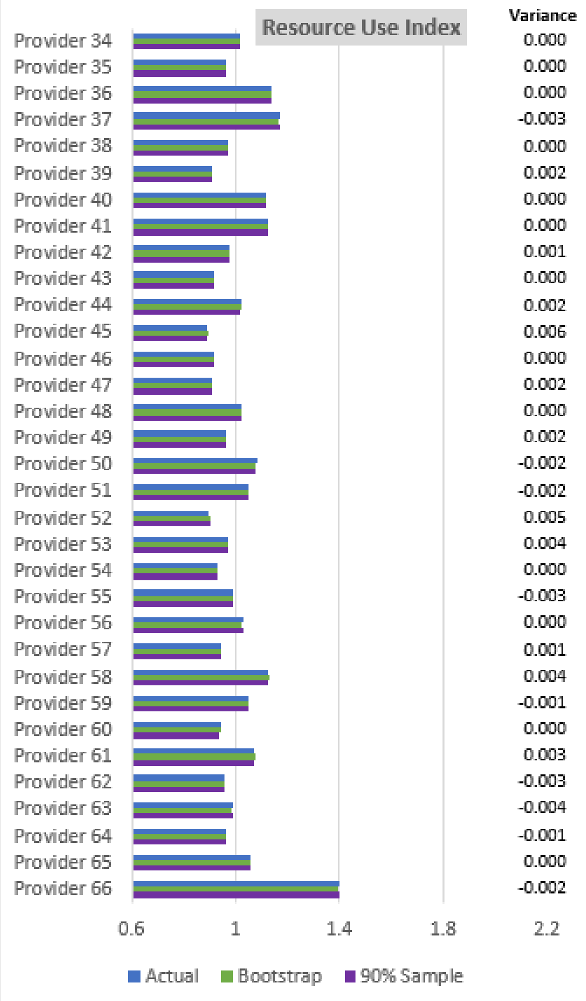
Reliability Paper describes the results of testing in detail:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188105.pdf

*Prior submission: Please see Appendix A (page 2) for prior submission results.*


Total Cost Index
Variance
Provider 01 -0.001
Provider 02 0.004
Provider 03 -0.001
Provider 04 0.001
Provider 05 -0.002
Provider 06 0.001
Provider 07 0.004
Provider 08 0.000
Provider 09 0.000
Provider 10 0.000
Provider 11 0.001
Provider 12 0.000
Provider 13 -0.001
Provider 14 0.000
Provider 15 -0.001
Provider 16 0.000
Provider 17 0.002
Provider 18 -0.001
Provider 19 0.000
Provider 20 0.000
Provider 21 0.001
Provider 22 0.001
Provider 23 0.001
Provider 24 -0.001
Provider 25 0.000
Provider 26 0.000
Provider 27 0.000
Provider 28 0.000
Provider 29 0.002
Provider 30 0.000
Provider 31 0.001
Provider 32 -0.003
Provider 33 0.003
0.6
1
1.4
1.8
2.2
Actual
Bootstrap
90% Sample


Resource Use Index
Variance
Provider 01 0.000
Provider 02 0.003
Provider 03 -0.001
Provider 04 0.002
Provider 05 -0.002
Provider 06 0.002
Provider 07 0.003
Provider 08 0.000
Provider 09 -0.001
Provider 10 0.000
Provider 11 0.002
Provider 12 0.000
Provider 13 0.000
Provider 14 0.001
Provider 15 -0.001
Provider 16 0.000
Provider 17 0.002
Provider 18 -0.001
Provider 19 0.000
Provider 20 0.000
Provider 21 0.001
Provider 22 0.001
Provider 23 0.001
Provider 24 -0.001
Provider 25 0.001
Provider 26 0.000
Provider 27 0.000
Provider 28 0.000
Provider 29 0.002
Provider 30 0.000
Provider 31 0.000
Provider 32 -0.001
Provider 33 0.002
0.6
1
1.4
1.8
2.2
Actual
Bootstrap
90% Sample


Total Cost Index
Variance
Provider 34
0.000
Provider 35
0.000
Provider 36
0.000
Provider 37
-0.003
Provider 38
0.000
Provider 39
0.003
Provider 40
0.000
Provider 41
-0.001
Provider 42
0.001
Provider 43
0.001
Provider 44
0.003
Provider 45
0.007
Provider 46
0.001
Provider 47
0.002
Provider 48
0.000
Provider 49
0.000
Provider 50
-0.002
Provider 51
0.000
Provider 52
0.008
Provider 53
0.004
Provider 54
0.001
Provider 55
-0.003
Provider 56
0.000
Provider 57
0.001
Provider 58
0.005
Provider 59
-0.001
Provider 60
-0.001
Provider 61
0.005
Provider 62
-0.004
Provider 63
-0.005
Provider 64
-0.002
Provider 65
0.002
Provider 66
-0.006
0.6
1
1.4
1.8
2.2
Actual
Bootstrap
90% Sample


Resource Use Index
Variance
Provider 34
0.000
Provider 35
0.000
Provider 36
0.000
Provider 37
-0.003
Provider 38
0.000
Provider 39
0.002
Provider 40
0.000
Provider 41
0.000
Provider 42
0.001
Provider 43
0.000
Provider 44
0.002
Provider 45
0.006
Provider 46
0.000
Provider 47
0.002
Provider 48
0.000
Provider 49
0.002
Provider 50
-0.002
Provider 51
-0.002
Provider 52
0.005
Provider 53
0.004
Provider 54
0.000
Provider 55
-0.003
Provider 56
0.000
Provider 57
0.001
Provider 58
0.004
Provider 59
-0.001
Provider 60
0.000
Provider 61
0.003
Provider 62
-0.003
Provider 63
-0.004
Provider 64
-0.001
Provider 65
0.000
Provider 66
-0.002
0.6
1
1.4
1.8
2.2
Actual
Bootstrap
90% Sample

| | Total Cost Index | | | | | Resource Use Index | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Provider Group | 90% Sample | Bootstrap | Actual | Variation between Actual and Bootstrap | Variation between Actual and 90% | 90% Sample | Bootstrap | Actual | Variation between Actual and Bootstrap | Variation between Actual and 90% |
| Provider 01 | 0.836 | 0.836 | 0.836 | (0.001) | (0.000) | 0.930 | 0.930 | 0.931 | (0.000) | (0.000) |
| Provider 02 | 0.841 | 0.845 | 0.842 | 0.004 | (0.001) | 0.951 | 0.955 | 0.951 | 0.003 | (0.001) |
| Provider 03 | 0.849 | 0.848 | 0.849 | (0.001) | (0.000) | 0.914 | 0.913 | 0.915 | (0.001) | (0.000) |
| Provider 04 | 0.868 | 0.869 | 0.868 | 0.001 | (0.000) | 0.968 | 0.970 | 0.968 | 0.002 | (0.000) |
| Provider 05 | 0.873 | 0.872 | 0.873 | (0.002) | (0.001) | 0.825 | 0.823 | 0.826 | (0.002) | (0.001) |
| Provider 06 | 0.883 | 0.885 | 0.884 | 0.001 | (0.001) | 0.960 | 0.963 | 0.961 | 0.002 | (0.001) |
| Provider 07 | 0.892 | 0.895 | 0.891 | 0.004 | 0.001 | 0.969 | 0.971 | 0.968 | 0.003 | 0.001 |
| Provider 08 | 0.902 | 0.903 | 0.903 | 0.000 | (0.000) | 0.940 | 0.941 | 0.940 | 0.000 | (0.000) |
| Provider 09 | 0.903 | 0.902 | 0.903 | (0.000) | 0.000 | 0.992 | 0.991 | 0.992 | (0.001) | 0.000 |
| Provider 10 | 0.904 | 0.904 | 0.904 | (0.000) | (0.000) | 0.981 | 0.981 | 0.981 | (0.000) | (0.000) |
| Provider 11 | 0.910 | 0.911 | 0.910 | 0.001 | 0.000 | 0.999 | 1.001 | 0.999 | 0.002 | 0.001 |
| Provider 12 | 0.911 | 0.911 | 0.911 | (0.000) | (0.000) | 0.980 | 0.980 | 0.980 | (0.000) | (0.000) |
| Provider 13 | 0.917 | 0.916 | 0.917 | (0.001) | (0.000) | 0.988 | 0.988 | 0.987 | 0.000 | 0.000 |
| Provider 14 | 0.918 | 0.918 | 0.917 | 0.000 | 0.000 | 0.947 | 0.947 | 0.946 | 0.001 | 0.000 |
| Provider 15 | 0.918 | 0.917 | 0.918 | (0.001) | (0.000) | 0.922 | 0.921 | 0.922 | (0.001) | (0.000) |
| Provider 16 | 0.926 | 0.926 | 0.926 | 0.000 | 0.000 | 1.019 | 1.020 | 1.019 | 0.000 | 0.000 |
| Provider 17 | 0.926 | 0.928 | 0.926 | 0.002 | (0.000) | 0.973 | 0.974 | 0.973 | 0.002 | (0.000) |
| Provider 18 | 0.945 | 0.943 | 0.944 | (0.001) | 0.000 | 0.894 | 0.892 | 0.893 | (0.001) | 0.000 |
| Provider 19 | 0.945 | 0.945 | 0.945 | (0.000) | (0.000) | 1.006 | 1.006 | 1.007 | (0.000) | (0.000) |
| Provider 20 | 0.957 | 0.957 | 0.958 | (0.000) | (0.000) | 0.981 | 0.981 | 0.981 | (0.000) | (0.000) |
| Provider 21 | 0.959 | 0.960 | 0.959 | 0.001 | 0.000 | 1.012 | 1.012 | 1.011 | 0.001 | 0.000 |
| Provider 22 | 0.960 | 0.962 | 0.960 | 0.001 | (0.000) | 0.869 | 0.871 | 0.870 | 0.001 | (0.001) |
| Provider 23 | 0.962 | 0.964 | 0.963 | 0.001 | (0.001) | 1.009 | 1.012 | 1.011 | 0.001 | (0.002) |
| Provider 24 | 0.974 | 0.973 | 0.973 | (0.001) | 0.000 | 1.033 | 1.032 | 1.032 | (0.001) | 0.000 |
| Provider 25 | 0.975 | 0.974 | 0.974 | (0.000) | 0.001 | 0.987 | 0.986 | 0.986 | 0.001 | 0.001 |
| Provider 26 | 0.976 | 0.976 | 0.976 | (0.000) | (0.000) | 0.997 | 0.997 | 0.997 | (0.000) | (0.000) |
| Provider 27 | 0.979 | 0.978 | 0.978 | (0.000) | 0.000 | 1.120 | 1.120 | 1.119 | 0.000 | 0.000 |
| Provider 28 | 0.985 | 0.985 | 0.985 | (0.000) | (0.000) | 1.041 | 1.040 | 1.041 | (0.000) | (0.000) |
| Provider 29 | 1.007 | 1.010 | 1.008 | 0.002 | (0.000) | 0.939 | 0.941 | 0.939 | 0.002 | (0.000) |
| Provider 30 | 1.014 | 1.013 | 1.013 | (0.000) | 0.001 | 1.024 | 1.022 | 1.022 | 0.000 | 0.002 |
| Provider 31 | 1.013 | 1.014 | 1.013 | 0.001 | (0.000) | 0.910 | 0.911 | 0.910 | 0.000 | 0.000 |
| Provider 32 | 1.019 | 1.016 | 1.019 | (0.003) | (0.000) | 1.042 | 1.040 | 1.042 | (0.001) | 0.000 |
| Provider 33 | 1.022 | 1.026 | 1.023 | 0.003 | (0.001) | 1.141 | 1.144 | 1.142 | 0.002 | (0.001) |
| Provider 34 | 1.026 | 1.026 | 1.026 | 0.000 | (0.000) | 1.017 | 1.017 | 1.017 | 0.000 | (0.000) |
| Provider 35 | 1.028 | 1.028 | 1.028 | (0.000) | (0.000) | 0.961 | 0.961 | 0.961 | 0.000 | (0.000) |
| Provider 36 | 1.038 | 1.037 | 1.037 | (0.000) | 0.000 | 1.140 | 1.139 | 1.139 | 0.000 | 0.000 |
| Provider 37 | 1.040 | 1.037 | 1.040 | (0.003) | (0.000) | 1.171 | 1.168 | 1.171 | (0.003) | (0.000) |
| Provider 38 | 1.052 | 1.051 | 1.051 | 0.000 | 0.000 | 0.968 | 0.967 | 0.968 | (0.000) | 0.000 |
| Provider 39 | 1.066 | 1.069 | 1.066 | 0.003 | 0.000 | 0.907 | 0.908 | 0.906 | 0.002 | 0.001 |
| Provider 40 | 1.066 | 1.066 | 1.066 | (0.000) | (0.000) | 1.116 | 1.116 | 1.116 | 0.000 | 0.000 |
| Provider 41 | 1.070 | 1.070 | 1.071 | (0.001) | (0.000) | 1.124 | 1.124 | 1.124 | (0.000) | (0.000) |
| Provider 42 | 1.074 | 1.075 | 1.074 | 0.001 | (0.001) | 0.978 | 0.979 | 0.979 | 0.001 | (0.000) |
| Provider 43 | 1.083 | 1.084 | 1.084 | 0.001 | (0.001) | 0.915 | 0.917 | 0.917 | 0.000 | (0.001) |
| Provider 44 | 1.100 | 1.104 | 1.101 | 0.003 | (0.001) | 1.020 | 1.023 | 1.021 | 0.002 | (0.001) |
| Provider 45 | 1.107 | 1.114 | 1.107 | 0.007 | 0.000 | 0.888 | 0.895 | 0.888 | 0.006 | (0.000) |
| Provider 46 | 1.112 | 1.113 | 1.112 | 0.001 | (0.000) | 0.916 | 0.916 | 0.916 | 0.000 | 0.000 |
| Provider 47 | 1.113 | 1.114 | 1.113 | 0.002 | 0.000 | 0.908 | 0.910 | 0.908 | 0.002 | 0.000 |
| Provider 48 | 1.117 | 1.118 | 1.118 | (0.000) | (0.001) | 1.022 | 1.023 | 1.022 | 0.000 | (0.001) |
| Provider 49 | 1.171 | 1.171 | 1.171 | 0.000 | (0.000) | 0.961 | 0.964 | 0.962 | 0.002 | (0.001) |
| Provider 50 | 1.180 | 1.179 | 1.182 | (0.002) | (0.002) | 1.081 | 1.080 | 1.082 | (0.002) | (0.001) |
| Provider 51 | 1.187 | 1.188 | 1.188 | (0.000) | (0.001) | 1.050 | 1.049 | 1.051 | (0.002) | (0.000) |
| Provider 52 | 1.191 | 1.199 | 1.191 | 0.008 | 0.000 | 0.899 | 0.904 | 0.899 | 0.005 | 0.000 |
| Provider 53 | 1.201 | 1.205 | 1.201 | 0.004 | (0.000) | 0.968 | 0.972 | 0.968 | 0.004 | (0.000) |
| Provider 54 | 1.203 | 1.203 | 1.203 | 0.001 | 0.001 | 0.927 | 0.927 | 0.926 | 0.000 | 0.001 |
| Provider 55 | 1.254 | 1.251 | 1.253 | (0.003) | 0.001 | 0.990 | 0.987 | 0.990 | (0.003) | 0.001 |
| Provider 56 | 1.255 | 1.255 | 1.255 | (0.000) | 0.000 | 1.028 | 1.027 | 1.028 | (0.000) | 0.000 |
| Provider 57 | 1.256 | 1.259 | 1.258 | 0.001 | (0.001) | 0.941 | 0.944 | 0.942 | 0.001 | (0.001) |
| Provider 58 | 1.266 | 1.274 | 1.268 | 0.005 | (0.002) | 1.123 | 1.129 | 1.125 | 0.004 | (0.002) |
| Provider 59 | 1.294 | 1.292 | 1.293 | (0.001) | 0.000 | 1.051 | 1.050 | 1.051 | (0.001) | 0.000 |
| Provider 60 | 1.359 | 1.359 | 1.359 | (0.001) | (0.000) | 0.940 | 0.940 | 0.940 | (0.000) | (0.000) |
| Provider 61 | 1.359 | 1.365 | 1.359 | 0.005 | (0.001) | 1.073 | 1.076 | 1.073 | 0.003 | (0.001) |
| Provider 62 | 1.423 | 1.420 | 1.424 | (0.004) | (0.001) | 0.958 | 0.956 | 0.959 | (0.003) | (0.001) |
| Provider 63 | 1.472 | 1.467 | 1.472 | (0.005) | (0.000) | 0.988 | 0.984 | 0.988 | (0.004) | 0.000 |
| Provider 64 | 1.538 | 1.535 | 1.538 | (0.002) | 0.000 | 0.965 | 0.964 | 0.965 | (0.001) | 0.000 |
| Provider 65 | 1.669 | 1.674 | 1.672 | 0.002 | (0.002) | 1.056 | 1.057 | 1.057 | (0.000) | (0.002) |
| Provider 66 | 2.027 | 2.022 | 2.028 | (0.006) | (0.000) | 1.398 | 1.396 | 1.399 | (0.002) | (0.000) |

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The results of the Bootstrap and Random Sample tests allow us to confidently conclude that the measures will reliably decipher TCI performance between levels of analysis (e.g. provider group).

- The bootstrap results indicate that the TCIs are reliable as the provider variation within all groups is <1% whereas the variation between groups spans >110%.

Reliability Paper describes the provider group results of testing in detail:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188105.pdf

______________________________

**2b2. VALIDITY TESTING**
**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)
x☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
x☐ **Performance measure score**
x☐ **Empirical validity testing**
x☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

*Prior submission: Please see Appendix A (page 14) for validity testing results from prior submission. The method of testing used for current resubmission is the same methodology used in prior submission.*

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

A Validity Analysis was performed on the HealthPartners' Total Cost of Care measure which indicates the results accurately reflect the performance levels of provider groups. The measure also accurately identifies the price (per unit cost) performance levels of providers. When used in conjunction with the Total Resource Use measure, the measure also accurately reflects resource use management across provider groups.

Detailed testing can be found in the Validity paper:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

Critical data elements

Non-risk adjusted correlations between ACG and Total Cost of Care, Total Cost Relative Resource Values (resource use) and utilization metrics were calculated.

Performance Measure Score

Risk adjusted Total Cost Index correlations to known risk adjusted utilization metrics were calculated.

Empirical testing of validity and overview of face validity policy and procedure

An assessment of high and low performing provider groups supports the relationship between risk adjusted utilization metrics and Total Cost Index.

The face validity process is conducted by transparently sharing results and methods with provider groups measured and allowing a 45-day comment period prior to public display of provider group results.

HealthPartners has a Policy and Procedure Review Process and executes it annually with each release of provider groups' performance and measurement results. Disclosure to providers includes:

1. Transparent reporting of measurement methodology
2. Providing comparative performance results with information on statistical reliability to providers
3. Providing an explanation of the results at least 45 days prior to their use in public reporting or business applications
4. Notifying providers of how the information will be used

5. A process by which providers can notify the plan of additional information or corrections

Public reporting of provider group measurement results:

https://www.healthpartners.com/public/cost-and-quality/index.html

Publicly available methods of rate calculations for transparency:

https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_033165.pdf

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

The correlation coefficients are included below for testing validity of the measure components and validity of the Total Cost of Care measure. Interpretation accompanies the tables of results below to provide context. However, please reference the paper to follow the complete analytical pathway with context and reasoning to conclude the measure is valid.
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

**Validity of Measure Components**

**Correlations Between ACG Score, Non-Risk Adjusted Per Member Per Month (PMPMs), Non-Risk Adjusted Total Cost Relative Resource Values (TCRRVs), and Risk Adjusted TCI**

| | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **ACG** | **Non-Risk Adj PMPMs** |
| Non-Risk Adj PMPM | 0.62 | 1.00 |
| Non-Risk Adj TCRRVs | 0.88 | 0.78 |
| ACG Risk Adj TCI | 0.03 | 0.79 |
| ACG Risk Adj RUI | 0.14 | 0.45 |
| Price | -0.09 | 0.57 |

- There is a high correlation between ACG score and the non-risk adjusted PMPM and TCRRVs which indicates that the non-risk adjusted PMPM and the non-risk adjusted TCRRVs are a good measure of resource use.
- There is a low correlation between ACG score and the risk adjusted TCI. This indicates that the risk score of a provider has no impact on a provider's ability to be a high performer.
- There is a low correlation between price and ACG because ACGs measure expected resource use whereas price is not affected by the number or intensity of services received. Price on the other hand is solely based on the provider and their referral partner's per unit cost and since overall costs are influenced by that per unit cost, price is highly correlated with non-risk adjusted PMPMs.

**Correlations Between the Non-Risk Adjusted Place of Service Metrics and Non-Risk Adjusted PMPMs & Non-Risk Adjusted TCRRVs**

| Non-Risk Adjusted | Correlation Coefficient | |
|---|---|---|
| *Service Category* Metric | *Non-Risk Adj Service Category PMPMs* | *Non-Risk Adj Service Category TCRRVs* |
| *Inpatient* | | |
| Admits/1000 | 0.67 | 0.82 |
| *Outpatient* | | |
| ER/1000 | 0.67 | 0.52 |
| OP Surgery/1000 | 0.60 | 0.68 |
| HighTech Rad/1000 | 0.45 | 0.67 |
| *Professional* | | |
| E&M/1000 | 0.63 | 0.71 |
| Lab/Path/1000 | 0.77 | 0.83 |
| Std Rad/1000 | 0.49 | 0.72 |
| *Pharmacy* | | |
| Rx/1000 | 0.73 | 0.80 |

**Inpatient:** There should be and are strong correlations between the admit rate to the non-risk adjusted PMPMs and non-risk adjusted TCRRVs as the only two factors not measured by the admits are the intensity and unit cost of the services performed.

**Outpatient:** There should be and are moderate correlations between the ER, outpatient surgery, and high tech radiology rates to the non-risk adjusted PMPMs and non-risk adjusted TCRRVs as these three utilization metrics combine to encompass approximately 65% of the total outpatient spend.

**Professional:** There should be and are  moderate correlations between the E&M visits, Lab/Path services, and standard radiology to the non-risk adjusted PMPMs and non-risk adjusted TCRRVs as they represent 45% of the professional spend, but are also good indicators of patients that consume medical services.

**Pharmacy:** There should be and are strong correlations between the pharmacy prescribing rates to the non-risk adjusted PMPMs and non-risk adjusted TCRRVs as the only factor that is not accounted for in the Rx prescribing rate metric is the intensity of the drug prescribed.  The intensity includes generic usage as well as the variation in cost between drugs.

Since the ACG score, non-risk adjusted PMPMs and non-risk adjusted TCRRVs are a measure of the consumption of health care services, there should be strong correlation between these values and known utilization metrics.

Composite Utilization:  A utilization metric was created by weighting each of the underlying utilization metrics by the place of service percent of resources it represents of the total resources by each provider group.

Composite Utilization Metric by Provider Group =
Inpatient (Admit Rate x Inpatient Resource Use %) +
Outpatient (Average (ER rate, OP Surg Rate, High Tech Rad Rate) x Outpatient Resource Use %) +
Professional (Average (E&M rate, Lab/Path Rate, Std Rad) x Professional Resource Use %) +
Pharmacy (Rx rate x Pharmacy Resource Use %)

| Non-Risk Adjusted | Correlation Coefficient | | |
|---|---|---|---|
| Metric | ACG | Non-Risk Adj PMPMs | Non-Risk Adj TCRRVs |
| Composite Utilization | 0.74 | 0.69 | 0.87 |

The non-risk adjusted resource composite is highly correlated with ACGs, non-risk adjusted PMPMs and non-risk adjusted TCRRVs.

## Validity of Total Cost of Care Measure

### Correlations Between the Risk Adjusted Place of Service Metrics and TCI, Price, and RUI

- Both overall Price and Total Resource Use are correlated with TCI as expected. However, price is more highly correlated with TCI as there is significantly more variation between providers in price than resource use, therefore it has a larger impact on TCI.
- Hospital-based care and professional TCI are strongly correlated with overall TCI.
- As expected both hospital-based care and professional price are strongly correlated with overall price.
- As expected there is little correlation between the Rx TCI and overall TCI as there is less variation in pharmacy when compared to the other places of service once ACG risk adjustment is applied.

| Risk Adjusted | Correlation Coefficient | | |
|---|---|---|---|
| **Metric** | **TCI** | **RUI** | **Price** |
| Hospital TCI | 0.74 | | |
| Prof TCI | 0.73 | | |
| Rx TCI | 0.16 | | |
| Hospital RUI | | 0.30 | |
| Prof RUI | | 0.74 | |
| Total RUI | 0.39 | | |
| Hospital Price | | | 0.86 |
| Prof Price | | | 0.83 |
| Total Price | 0.87 | | |

### Correlations Between Risk Adjusted Place of Service Utilization Metrics and Corresponding TCI

| Risk Adjusted | Correlation Coefficient | |
|---|---|---|
| ***Service Category* Metric** | ***Service Category TCIs*** | ***Service Category RUIs*** |
| ***Inpatient*** | | |
| Admit Rate | 0.78 | 0.82 |
| ***Outpatient*** | | |
| ER Cnt | 0.68 | 0.46 |
| OP Surgery | 0.55 | 0.49 |
| High Tech Rad | 0.21 | 0.37 |
| ***Professional*** | | |
| E&M Visits | 0.48 | 0.70 |
| Lab/Path | 0.59 | 0.54 |
| Std Rad | 0.48 | 0.38 |
| ***Pharmacy*** | | |
| Rx Count | 0.25 | |

**Inpatient:** There is a high correlation between the risk adjusted admit rate and the inpatient TCI. This would indicate that the higher the risk adjusted admit rate the more likely a provider will have a higher than average TCI.

**Outpatient:** There is a moderate correlation between the risk adjusted ER count and the outpatient TCI. This would indicate that the higher the risk adjusted ER counts the more likely a provider will have a higher than average outpatient TCI.

**Professional:** The professional utilization metrics are moderately correlated to the professional TCI.

This result is as expected because the professional place of service includes a significant amount of services beyond these three utilization measures (other professional services = 55%).

It is also as expected because having higher than average utilization on diagnostic or management based services does not necessarily indicate a higher resource consuming patient.

**Pharmacy:** The low correlation between Rx count and Rx TCI indicates that after risk adjustment the type and cost of the drug prescribed (e.g., brand vs generic) drives TCI rather than the number of prescriptions.

| Risk Adjusted | Correlation Coefficient | Correlation Coefficient |
|---|---|---|
| **Metric** | **TCI** | **RUI** |
| Composite Utilization | 0.72 | 0.52 |

The indexed Total Cost of Care measure has a high correlation to a risk adjusted composite utilization index, which was developed as a proxy to measure total resource consumption.

*Prior submission: Please see Appendix A (page 14) for prior submission results.*

In addition, the Total Cost of Care measure was analyzed over time (2013 through 2015) to demonstrate stability and sensitivity to provider changes or improvement initiatives. Providers' performance across all three measures is relatively consistent across all three years and results are shown in the table below. The factors that drive variation between years within a provider are cost per unit and resource use management.

The results show that TCI has the most variation as it combines the changes for both price and resource use. The results also show that there is more variation in resource use over time than price. This indicates that providers are receiving similar price increases, but how providers are managing their patients' resource use is contributing more to the variation seen in costs.

| | TCI | | | | Price | | | | RUI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Provider Group Size** | **25th Percentile** | **Average** | **Median** | **75th Percentile** | **25th Percentile** | **Average** | **Median** | **75th Percentile** | **25th Percentile** | **Average** | **Median** | **75th Percentile** |
| <1,000 | 0.04 | 0.07 | 0.07 | 0.11 | 0.02 | 0.04 | 0.03 | 0.05 | 0.03 | 0.05 | 0.05 | 0.09 |
| 1,000-2,000 | 0.03 | 0.08 | 0.07 | 0.11 | 0.02 | 0.03 | 0.03 | 0.04 | 0.03 | 0.06 | 0.07 | 0.09 |
| 2,000+ | 0.01 | 0.03 | 0.03 | 0.04 | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 | 0.03 | 0.03 | 0.05 |

*Prior submission: Please see Appendix A (page 6) for prior submission results.*

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The Total Cost of Care measure is valid as the critical data elements and the criteria applied produce a measure that accurately assesses various levels of performance. The norms in the measure are the network averages from the healthcare information derived from the MN market from included entities.

The Validity paper describes the results and conclusions from testing in detail:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

In summary, the Total Cost of Care measure accurately and consistently identified providers that are low or high performers with conclusions supported by known utilization measures.

There are high correlations between non-risk adjusted PMPM, ACG score and non-risk adjusted TCRRVs which indicate they are good measures of resources.

The ACGs, non-risk adjusted PMPMs, and non-risk adjusted TCRRVs have similar correlations to all utilization metrics which indicates the TCRRVs are performing as expected and are a solid measure of resources.

Both overall Price and Total Resource Use are highly correlated with TCI as expected.

The indexed Total Cost of Care measure scores have a high correlation (0.72) to a risk adjusted composite utilization index score, which was developed as a proxy to measure total resource consumption.

The Total Cost of Care measure differentiates between provider groups accurately as supported by the risk adjusted service utilization metrics, resource use and price measures.

___________________________

**2b3. EXCLUSIONS ANALYSIS**

**NA □ no exclusions — *skip to section 2b4***

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
The HealthPartners' Total Cost of Care measure is a full population-based measure, with members under age 1, members 65+ and members with less than 9 months of enrollment excluded to ensure an accurate risk assessment is made on the population.

- Members over age 64
- Members under age 1
- Member enrollment less than nine months during the one year measurement time window
- Dollars per member up to $125,000 are included; dollars per member above $125,000 are excluded (truncated)

*Prior submission: For this maintenance submission, the only change to HealthPartners Total Cost of Care measure from prior submission is the truncation level. The total spend truncation level for a member's combined medical and pharmacy claims has increased from $100,000 to $125,000 to account for the natural rise in healthcare costs over the past several years.*

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Results from testing truncation level at $125,000 can be found in the Validity paper. No other changes to measure criteria have occurred since endorsement.
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

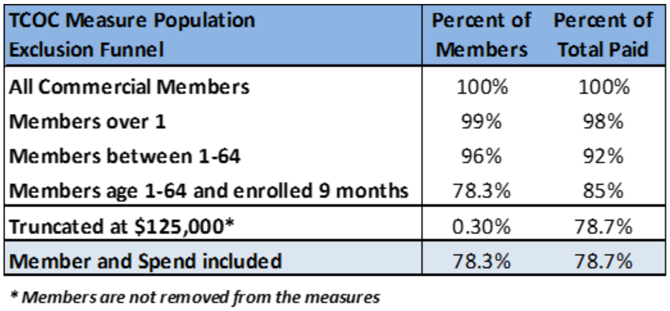
The truncation was increased from $100,000 to $125,000. Given medical inflation has been 2-4% per year recently, it is necessary to increase the spend truncation to account for the natural rise in healthcare costs. Since the model needs to remain stable year over year, the truncation level also needs to remain stable, with only periodic updates. The $125,000 truncation level returns the model to its original NQF endorsed state in terms of R-squared, percent of dollars included in the model.

The following exclusions and decision points remain unchanged from the original endorsed measures.

Nine month continuous enrollment – A nine month continuous enrollment was selected to balance business operations. Nine months allows for partial year enrollee. There was very little statistical difference in R-squared between six and twelve months.

Infants, under age one are excluded due to slightly higher R-squared of the population without newborns, the required nine months enrollment criteria and variability in newborn costs, newborns under age one were excluded from the total cost of care measure.

Members over age 64 due are excluded due to potential incomplete claims data of Medicare eligible beneficiary.



| TCOC Measure Population Exclusion Funnel | Percent of Members | Percent of Total Paid |
|---|---|---|
| All Commercial Members | 100% | 100% |
| Members over 1 | 99% | 98% |
| Members between 1-64 | 96% | 92% |
| Members age 1-64 and enrolled 9 months | 78.3% | 85% |
| Truncated at $125,000* | 0.30% | 78.7% |
| Member and Spend included | 78.3% | 78.7% |

** Members are not removed from the measures*

________________________________

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.***

**2b4.1. What method of controlling for differences in case mix is used?**

☐ **No risk adjustment or stratification**

☐ **Statistical risk model with** 0 **risk factors**

☐ **Stratification by** 0 **risk categories**

x☐ **Other, Johns Hopkins ACG System on commercially covered population**

**2b4.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

The Total Cost of Care measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). The measure is also limited by insurance coverage to commercial only.

The ACG System is a statistically valid and broadly adopted risk grouper in both academic and non-academic settings with methodology derived from diagnosis information. Information about the development of the grouper can be found here: http://acg.jhsph.org/; additionally please refer to the ACG Technical Reference Guide for supporting material:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf

ACG Grouper:

- Adjusted Clinical Groups (ACG System) were developed by Johns Hopkins University and allow comparisons between populations with varying illness burdens based on diagnoses, age and gender.
- Each unique member is assigned one of 93 ACG actuarial cells, which has a corresponding weight that reflects relative illness burden (e.g. relative expected resource consumption). Attributed members are assigned a risk score based on diagnoses on claims from the performance measurement period, as well as member age and gender

ACG-cell Risk Weights/Coefficients:

- The ACG risk weights measure relative resource variation between ACG actuarial cells/codes. Please see page 30-34 of the reference guide to view each ACG-cell risk weight. https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf
- Multiply each member's ACG weight by their eligible member months creating a total member ACG weight.

ACG Score:

- Each provider's attributed member ACG weights are summed to the provider level and divided by the sum of the attributed member months creating an ACG score for the provider.
- The provider's average ACG score is indexed to all attributed member's plan average ACG score.

- A member's total member ACG weight is updated to correspond with each year the Total Cost of Care measure is measured.

Each of the 93 ACG actuarial cells can be considered a covariate of the multivariate risk model with the cell weights being the coefficients. The ACG cells are non-linear composites of the three risk factors: age, gender, diagnosis. Each member is assigned one of 93 covariates in the multivariate model and is based on the member's combination of age, gender and complete history of diagnosis codes.

**2b4.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

Not applicable. All measures are clinically risk adjusted and limited to the commercial population.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

The Total Cost of Care measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). The measure is also limited by insurance coverage to commercial only.

The ACG System is a statistically valid and broadly adopted risk grouper in both academic and non-academic settings with methodology derived from diagnosis information. Information about the development of the grouper can be found here: http://acg.jhsph.org/; additionally please refer to the ACG Technical Reference Guide for supporting material:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf

The ACG System assigns International Classification of Disease (ICD) diagnosis codes to 32 diagnosis groups – Aggregated Diagnosis Groups (ADGs). The assignment method is included in the ACG software for all codes. Diagnosis codes mapped to a given ADG are clinically similar and have similar expected need for healthcare resources. The assignment criteria is based on features of a condition that help predict duration and intensity of resource use. Five clinical criteria are used to determine assignment of codes: duration, severity, diagnostic certainty, type of etiology, and expected need for specialty care. The 32 ADGs are listed on pages 4-6 in the reference guide, along with a table on pages 8-10 that provides guidance on how the five criteria are applied to each ADG.

Adjusted Clinical Group actuarial cells (ACGs) build off of the ADG assignment logic described and are used to determine the morbidity profile of patient populations to more fairly assess provider performance and allow for equitable comparisons of utilization and outcomes. ACGs are defined by morbidity, age, and sex and are person-focused to categorize patients' illnesses. Based on the pattern of morbidities, the ACG approach assigns each individual to a single ACG category. The ACG assignment process can be found on page 12 of the reference guide.

After applying measure criteria, which includes limitation to commercial only and clinical risk adjustment, socioeconomic testing was conducted that considered income and education status as potential factors beyond those already adjusted for. Methodology and testing results can be found here:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

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

The risk factors included in ACG risk grouper were determined in the development of Johns Hopkins ACG risk grouper and are not available to the general public. The performance of the risk groupers are the basis for verifying the risk factors included in the model are sufficient to address clinic risk variation. The Society of Actuaries Accuracy of Claims-Based Risk Scoring Models (2016) findings also indicate the reliability and validity of the ACG risk grouper. https://www.soa.org/Files/Research/research-2016-accuracy-claims-based-risk-scoring-models.pdf

**2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)**

After risk adjusting for age, gender, and clinical risk, and limiting by insurance type, income does not significantly impact a patient's total cost. As a potential practical use case example, the study also evaluated Resource Use provider group performance and found there was no discernible difference in performance when adjusting for income. The provider group analysis focused on the Resource Use measure to remove any bias based on price. The study considered two different data sources to study income variation, Census tract data and a commercially licensed data source available to HealthPartners with more specific income data.

https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

The study utilized two independent data sources to evaluate income. The first was U.S. Census tracts. As defined by the U.S. Department of Commerce, "Census tracts are small, relatively permanent statistical subdivisions of a county or equivalent entity that are updated by local participants prior to each decennial census as part of the Census Bureau's Participant Statistical Areas Program. The Census Bureau delineates census tracts in situations where no local participant existed or where state, local, or tribal governments declined to participate. The primary purpose of census tracts is to provide a stable set of geographic units for the presentation of statistical data.

Census tracts generally have a population size between 1,200 and 8,000 people, with an optimum size of 4,000 people. A census tract usually covers a contiguous area; however, the spatial size of census tracts varies widely depending on the density of settlement. Census tract boundaries are delineated with the intention of being maintained over a long time so that statistical comparisons can be made from census to census. Census tracts occasionally are split due to population growth or merged as a result of substantial population decline."[5] As noted, tracts estimate income by a general area and are not highly specific, introducing potential error and bias in the model.

HealthPartners utilized an additional data source to more accurately assess household income for purposes of this study. HealthPartners commercially licenses and has access to a large consumer database for other business purposes which gave us the ability to evaluate income with more specificity at the household level. Recognizing that it may not be feasible for all users to access a commercial database, HealthPartners pursued this deeper evaluation to more broadly understand the important question of whether or not to adjust cost performance measures by socioeconomic status independent of data availability. Household level income is derived using the midpoint of defined ranges of income by household (e.g. $20,000-$30,000) and capped at $250,000. Using the midpoint of a range introduces potential error in the evaluation whereas self-reported individual or household income would be most accurate.

**Population-Based Evaluation**

The evaluation tested the inclusion of income in addition to the factors already included in the measure specifications - age, gender, and clinical risk. Detailed measure criteria can be found in HealthPartners Technical Guidelines.

The study population included HealthPartners' full book of business of members, Commercial and Medicaid with TCOC criteria applied using services and claims generated throughout the 2015 time period. The study population included more than 530,000 members.

Three multiple linear regression models were created, each with one of the three metrics of interest as the dependent variable (total reimbursed amount per member per month, resource use per member per month, and price).  Each model was identical in the use of income, ACG risk score, and insurance product (commercial vs Medicaid) as the independent variables. Resource use, reimbursed amount, price, and ACG scores were log transformed prior to developing the regression models to address the skewed nature of the data and adjust for heteroscedasticity. Insurance product was treated as a binary variable (commercial = 1, Medicaid = 0). The resulting coefficients were analyzed in terms of a 1% increase from average and their corresponding effect on the dependent variables.

Additionally, a model was created using only the endorsed measure criteria for the Resource Use measure (i.e. ACG and product only as the independent variables). The $R^2$ statistic from this model was compared against the $R^2$ statistic from the model that included income as an independent variable, allowing us to quantify the predictive value of income on resource use.

The same regression statistics and models were used with the second, more robust data source available to HealthPartners. This data contained more accurate income information which was specific to household rather than tract, with household income defined using the midpoint or median of the income ranges. The more robust data source was available for 65% of HealthPartners' book of business members for 2015 and in the same proportions of commercial to Medicaid as in the previous evaluation.

**Provider Group Performance Evaluation**

A second evaluation was performed to provide a potential practical example of adjusting the TCOC and resource use measures by income using the Census and commercially licensed data sources. Resource Use Index was evaluated to remove known price variations between providers. HealthPartners' resource use results for its primary care network of commercial attributed members were used to evaluate provider group performance when adjusting for income. Medicaid was excluded from this evaluation as it has already been determined that provider performance results should be segmented by product.

There were 66 provider groups who met the measure criteria and were included in the evaluation using the Census tract data. The TCOC measure is endorsed at a reliability level of 600 patients. Because the commercially licensed data source had available data for 65% of HealthPartners' book of business, there were 11 provider groups that failed to meet the 600 minimum and were excluded from the evaluation.

The variation between the average incomes using the Census tract data or the commercially licensed data source for each provider group was compared to the network average to adjust the provider's resource use index. It should be noted that while the adjustment can be made, the results should not be considered valid or reliable given the limitations inherent in each data source as described previously.

The regression analysis generated parameters that were translated into results based upon average cost, resource use, income, and ACG scores.

**Table of Regression results using Census Tract Data**

| Model | 1% Income Increase | 1% ACG increase | Commercial vs. Medicaid Membership |
|---|---|---|---|
| Total Reimbursement | $ (0.13) | $ 4.22 | $ 133.28 |
| Resource Use | $ 0.16 | $ 4.34 | $ (75.24) |
| Price | $ (0.28) | $ 0.07 | $ 205.36 |

| MODEL | R_SQUARED |
|---|---|
| Resource Use Endorsed Measure | 0.5788 |
| Resource Use Endorsed Measure + Income | 0.5792 |

Using Census tract data, a 1% increase in income resulted in a $0.13 decrease in total reimbursement, a $0.16 increase in resource use, and $0.28 decrease in price. The results highlight how significantly more the ACG score (clinical risk adjustment) and insurance product impact both the cost and resource use measures. For frame of reference, on average for the Midwest market, the total spend for a member per month (PMPM) is $400. The results of the evaluation show that a 1% increase in risk score accounts for a $4.22 or roughly 1% increase in PMPM.

Product also contributed significantly with there being a $133 dollar difference in cost between commercial and Medicaid. The variation in resource use was much less, however, still significant with Medicaid covered members utilizing $75 more dollars of resources. The fact that Medicaid's cost per service is approximately half that of commercial rates drives the differences between the TCOC and Resource Use results. The R$^2$ results further emphasize that ACG score and insurance type are the main drivers of cost and resource use variation and income does not provide any additional predictive power.

**Table of Regression results using Commercially Licensed Data Source**

| Model | 1% Income Increase | 1% ACG increase | Commercial vs. Medicaid Membership |
|---|---|---|---|
| Total Reimbursement | $ (0.00) | $ 4.56 | $ 139.80 |
| Resource Use | $ 0.05 | $ 4.66 | $ (81.26) |
| Price | $ (0.07) | $ 0.06 | $ 218.13 |

| MODEL | R_SQUARED |
|---|---|
| Resource Use Endorsed Measure | 0.57318 |
| Resource Use Endorsed Measure + Income | 0.57321 |

Using the commercially purchased data source, with income by household, a 1% increase in income resulted in no change for total reimbursement, $0.05 increase in resource use, and $0.07 decrease in price. This is telling, as when using a data source that is more specific, income is even less impactful on TCOC and resource use while ACG and product type show similar results.

**Results– Provider Group Performance Evaluation**

Provider group performance of the Resource Use measure was evaluated to test the impact of income adjustment on the Resource Use measure. Provider group results for both data sources, Census tract and commercially licensed, are shown below using HealthPartners' commercial provider network. The Resource Use Index (RUI) is calculated using the endorsed measure criteria. The second RUI is calculated using the endorsed measure criteria *with income adjustment.*

The Census tract data evaluated 66 provider groups and the commercially licensed data source evaluated 55 provider groups. Because the population of patients used between the two data sources is different, Provider Group 01 in the Census tract chart is not the same as Provider Group 01 in the commercially licensed chart. Provider group numbers in the Census tract chart are numbered based on ascending Total Cost Index found in the appendix of the study paper. Provider groups for both charts are sorted in ascending order using the RUI.

On average there was less than a 1% change in performance for provider groups when income was introduced into the model for the Resource Use measure when using Census tract data. This impact was reduced on average to less than a 0.25% when using the commercially licensed data source with more specific income data. Considering the Resource Use measure identifies provider performance levels (indices) that span greater than 167% as identified below, the less than 1% adjustment was considered insignificant when comparing provider performance. Provider Group charts begin on the following page.

**Census Tract Data Source**

| | |
|---|---|
| RUI Min | 0.82 |
| RUI Max | 1.39 |
| RUI Max/Min % Difference | 167% |
| Average % change with income adjustment | 0.64% |

**Commercially Licensed Data Source**

| | |
|---|---|
| RUI Min | 0.83 |
| RUI Max | 1.39 |
| RUI Max/Min % Difference | 167% |
| Average % change with income adjustment | 0.19% |

**Provider Group Detailed Results – Census Data**

| Provider Group | TCI | RUI (endorsed measure) | RUI (endorsed measure + income) | Average Income | RUI Change | Pct RUI Change |
|---|---|---|---|---|---|---|
| Provider 05 | 0.87 | 0.82 | 0.83 | $51,182.66 | 0.0097 | 1.18% |
| Provider 22 | 0.96 | 0.87 | 0.87 | $60,871.13 | 0.0051 | 0.59% |
| Provider 45 | 1.11 | 0.88 | 0.89 | $51,196.01 | 0.0097 | 1.10% |
| Provider 52 | 1.19 | 0.89 | 0.90 | $57,184.20 | 0.0069 | 0.77% |
| Provider 18 | 0.94 | 0.89 | 0.90 | $53,262.19 | 0.0087 | 0.98% |
| Provider 39 | 1.07 | 0.90 | 0.91 | $52,994.97 | 0.0089 | 0.98% |
| Provider 47 | 1.11 | 0.90 | 0.92 | $48,573.69 | 0.0110 | 1.21% |
| Provider 31 | 1.01 | 0.91 | 0.91 | $54,522.47 | 0.0081 | 0.90% |
| Provider 46 | 1.11 | 0.91 | 0.92 | $55,143.95 | 0.0079 | 0.86% |
| Provider 43 | 1.08 | 0.91 | 0.93 | $49,821.34 | 0.0104 | 1.13% |
| Provider 03 | 0.85 | 0.92 | 0.91 | $74,230.68 | -0.0012 | -0.13% |
| Provider 15 | 0.92 | 0.92 | 0.93 | $54,236.28 | 0.0083 | 0.90% |
| Provider 54 | 1.20 | 0.92 | 0.93 | $59,432.45 | 0.0058 | 0.63% |
| Provider 60 | 1.36 | 0.93 | 0.93 | $57,038.75 | 0.0070 | 0.75% |
| Provider 01 | 0.84 | 0.93 | 0.94 | $59,923.30 | 0.0056 | 0.60% |
| Provider 29 | 1.01 | 0.94 | 0.94 | $56,657.27 | 0.0071 | 0.76% |
| Provider 57 | 1.26 | 0.94 | 0.95 | $46,884.50 | 0.0117 | 1.25% |
| Provider 08 | 0.90 | 0.94 | 0.94 | $74,671.53 | -0.0014 | -0.14% |
| Provider 64 | 1.54 | 0.95 | 0.96 | $50,511.60 | 0.0100 | 1.06% |
| Provider 62 | 1.42 | 0.95 | 0.96 | $51,481.13 | 0.0096 | 1.01% |
| Provider 49 | 1.17 | 0.95 | 0.95 | $63,017.62 | 0.0041 | 0.44% |
| Provider 14 | 0.92 | 0.95 | 0.94 | $85,046.08 | -0.0063 | -0.66% |
| Provider 02 | 0.84 | 0.96 | 0.96 | $75,988.94 | -0.0020 | -0.21% |
| Provider 35 | 1.03 | 0.96 | 0.97 | $53,580.68 | 0.0086 | 0.89% |
| Provider 53 | 1.20 | 0.96 | 0.97 | $60,513.25 | 0.0053 | 0.55% |
| Provider 42 | 1.07 | 0.96 | 0.97 | $56,581.35 | 0.0072 | 0.74% |
| Provider 38 | 1.05 | 0.96 | 0.97 | $53,033.63 | 0.0088 | 0.92% |
| Provider 06 | 0.88 | 0.96 | 0.96 | $78,737.12 | -0.0033 | -0.34% |
| Provider 63 | 1.47 | 0.97 | 0.97 | $68,995.93 | 0.0013 | 0.14% |
| Provider 04 | 0.87 | 0.97 | 0.97 | $63,162.88 | 0.0041 | 0.42% |
| Provider 07 | 0.89 | 0.97 | 0.96 | $87,449.16 | -0.0074 | -0.76% |
| Provider 17 | 0.93 | 0.97 | 0.97 | $75,724.77 | -0.0019 | -0.19% |
| Provider 20 | 0.96 | 0.98 | 0.98 | $81,800.09 | -0.0047 | -0.48% |

**Provider Group Detailed Results – Census Data - *continued***

| Provider Group | TCI | RUI (endorsed measure) | RUI (endorsed measure + income) | Average Income | RUI Change | Pct RUI Change |
|---|---|---|---|---|---|---|
| Provider 10 | 0.90 | 0.98 | 0.98 | $80,344.27 | -0.0040 | -0.41% |
| Provider 12 | 0.91 | 0.98 | 0.99 | $68,200.83 | 0.0017 | 0.17% |
| Provider 55 | 1.25 | 0.98 | 1.00 | $45,478.38 | 0.0124 | 1.26% |
| Provider 25 | 0.97 | 0.99 | 0.99 | $55,181.82 | 0.0078 | 0.79% |
| Provider 13 | 0.92 | 0.99 | 0.98 | $85,397.81 | -0.0064 | -0.65% |
| Provider 09 | 0.90 | 1.00 | 0.99 | $79,078.76 | -0.0034 | -0.35% |
| Provider 26 | 0.98 | 1.00 | 1.00 | $76,886.96 | -0.0024 | -0.24% |
| Provider 11 | 0.91 | 1.00 | 1.01 | $58,557.65 | 0.0062 | 0.62% |
| Provider 19 | 0.94 | 1.01 | 1.01 | $76,364.65 | -0.0022 | -0.21% |
| Provider 23 | 0.96 | 1.01 | 1.02 | $51,695.82 | 0.0095 | 0.94% |
| Provider 21 | 0.96 | 1.01 | 1.01 | $80,133.18 | -0.0039 | -0.39% |
| Provider 48 | 1.12 | 1.02 | 1.02 | $62,718.98 | 0.0043 | 0.42% |
| Provider 34 | 1.03 | 1.02 | 1.01 | $76,650.16 | -0.0023 | -0.23% |
| Provider 44 | 1.10 | 1.02 | 1.02 | $57,718.34 | 0.0066 | 0.65% |
| Provider 30 | 1.01 | 1.02 | 1.03 | $60,952.95 | 0.0051 | 0.50% |
| Provider 56 | 1.26 | 1.02 | 1.03 | $56,343.84 | 0.0073 | 0.71% |
| Provider 16 | 0.93 | 1.03 | 1.02 | $73,585.43 | -0.0009 | -0.08% |
| Provider 24 | 0.97 | 1.03 | 1.04 | $61,287.61 | 0.0050 | 0.48% |
| Provider 32 | 1.02 | 1.04 | 1.03 | $88,286.87 | -0.0078 | -0.75% |
| Provider 28 | 0.98 | 1.05 | 1.04 | $76,082.19 | -0.0020 | -0.19% |
| Provider 51 | 1.19 | 1.05 | 1.04 | $80,419.35 | -0.0041 | -0.39% |
| Provider 59 | 1.29 | 1.05 | 1.06 | $55,164.25 | 0.0078 | 0.75% |
| Provider 65 | 1.67 | 1.05 | 1.06 | $55,820.84 | 0.0075 | 0.72% |
| Provider 61 | 1.36 | 1.06 | 1.07 | $60,338.84 | 0.0054 | 0.51% |
| Provider 50 | 1.18 | 1.08 | 1.09 | $42,557.01 | 0.0138 | 1.28% |
| Provider 40 | 1.07 | 1.12 | 1.11 | $94,343.76 | -0.0106 | -0.95% |
| Provider 58 | 1.27 | 1.12 | 1.13 | $52,722.55 | 0.0090 | 0.80% |
| Provider 27 | 0.98 | 1.12 | 1.12 | $85,490.55 | -0.0065 | -0.58% |
| Provider 41 | 1.07 | 1.13 | 1.12 | $86,685.54 | -0.0070 | -0.62% |
| Provider 36 | 1.04 | 1.14 | 1.14 | $82,723.92 | -0.0052 | -0.45% |
| Provider 33 | 1.02 | 1.15 | 1.14 | $89,318.28 | -0.0083 | -0.72% |
| Provider 37 | 1.04 | 1.18 | 1.17 | $85,086.95 | -0.0063 | -0.53% |
| Provider 66 | 2.03 | 1.39 | 1.38 | $75,167.49 | -0.0016 | -0.12% |

## Provider Group Detailed Results - Commercially Licensed Data

| Provider Group | TCI | RUI (endorsed measure) | RUI (endorsed measure + income) | Average Income | RUI Change | Pct RUI Change |
|---|---|---|---|---|---|---|
| Provider 01 | 0.93 | 0.83 | 0.84 | $ 72,443.51 | 0.0029 | 0.34% |
| Provider 02 | 1.08 | 0.88 | 0.88 | $ 81,736.10 | 0.0017 | 0.19% |
| Provider 03 | 1.14 | 0.88 | 0.89 | $ 77,260.46 | 0.0022 | 0.25% |
| Provider 04 | 1.03 | 0.89 | 0.89 | $ 80,589.91 | 0.0018 | 0.20% |
| Provider 05 | 1.00 | 0.90 | 0.90 | $ 79,052.16 | 0.0020 | 0.22% |
| Provider 06 | 0.85 | 0.90 | 0.90 | $ 90,000.35 | 0.0006 | 0.07% |
| Provider 07 | 0.86 | 0.92 | 0.92 | $ 81,080.58 | 0.0018 | 0.19% |
| Provider 08 | 1.01 | 0.92 | 0.92 | $ 79,498.09 | 0.0020 | 0.21% |
| Provider 09 | 1.41 | 0.92 | 0.93 | $ 81,478.38 | 0.0017 | 0.18% |
| Provider 10 | 1.08 | 0.93 | 0.93 | $ 75,610.49 | 0.0025 | 0.27% |
| Provider 11 | 0.97 | 0.93 | 0.94 | $ 79,045.36 | 0.0020 | 0.22% |
| Provider 12 | 1.62 | 0.94 | 0.94 | $ 75,077.69 | 0.0025 | 0.27% |
| Provider 13 | 1.21 | 0.94 | 0.94 | $ 83,735.62 | 0.0014 | 0.15% |
| Provider 14 | 0.93 | 0.95 | 0.94 | $ 95,896.16 | -0.0002 | -0.02% |
| Provider 15 | 1.19 | 0.95 | 0.95 | $ 82,285.13 | 0.0016 | 0.17% |
| Provider 16 | 0.92 | 0.96 | 0.96 | $ 75,238.24 | 0.0025 | 0.26% |
| Provider 17 | 1.10 | 0.96 | 0.96 | $ 79,490.54 | 0.0020 | 0.20% |
| Provider 18 | 0.95 | 0.96 | 0.96 | $102,194.19 | -0.0010 | -0.10% |
| Provider 19 | 0.85 | 0.97 | 0.97 | $ 74,225.15 | 0.0026 | 0.27% |
| Provider 20 | 1.54 | 0.97 | 0.97 | $ 80,176.59 | 0.0019 | 0.19% |
| Provider 21 | 1.10 | 0.98 | 0.98 | $ 76,567.93 | 0.0023 | 0.24% |
| Provider 22 | 0.93 | 0.98 | 0.98 | $105,426.41 | -0.0014 | -0.14% |
| Provider 23 | 1.32 | 0.98 | 0.99 | $ 72,760.80 | 0.0028 | 0.29% |
| Provider 24 | 0.96 | 0.98 | 0.98 | $114,650.89 | -0.0026 | -0.26% |
| Provider 25 | 0.89 | 0.98 | 0.99 | $ 87,932.23 | 0.0009 | 0.09% |
| Provider 26 | 0.95 | 0.99 | 0.98 | $107,107.13 | -0.0016 | -0.16% |
| Provider 27 | 0.92 | 0.99 | 0.99 | $ 83,708.65 | 0.0014 | 0.14% |

**Provider Group Detailed Results - Commercially Licensed Data - *continued***

| Provider Group | TCI | RUI (endorsed measure) | RUI (endorsed measure + income) | Average Income | RUI Change | Pct RUI Change |
|---|---|---|---|---|---|---|
| Provider 28 | 0.95 | 0.99 | 0.99 | $100,567.76 | -0.0008 | -0.08% |
| Provider 29 | 0.95 | 0.99 | 0.99 | $109,144.48 | -0.0019 | -0.19% |
| Provider 30 | 0.98 | 0.99 | 0.99 | $105,829.78 | -0.0014 | -0.14% |
| Provider 31 | 1.00 | 0.99 | 0.99 | $ 98,557.24 | -0.0005 | -0.05% |
| Provider 32 | 0.92 | 0.99 | 0.99 | $106,951.43 | -0.0016 | -0.16% |
| Provider 33 | 1.00 | 1.00 | 0.99 | $108,508.94 | -0.0018 | -0.18% |
| Provider 34 | 1.13 | 1.00 | 1.00 | $ 77,921.25 | 0.0022 | 0.22% |
| Provider 35 | 0.91 | 1.00 | 1.00 | $ 99,877.19 | -0.0007 | -0.07% |
| Provider 36 | 1.00 | 1.01 | 1.01 | $ 74,293.07 | 0.0026 | 0.26% |
| Provider 37 | 1.31 | 1.01 | 1.02 | $ 82,705.61 | 0.0015 | 0.15% |
| Provider 38 | 1.58 | 1.01 | 1.02 | $ 88,328.20 | 0.0008 | 0.08% |
| Provider 39 | 1.04 | 1.02 | 1.02 | $100,477.23 | -0.0007 | -0.07% |
| Provider 40 | 1.00 | 1.02 | 1.02 | $ 83,196.31 | 0.0015 | 0.15% |
| Provider 41 | 1.19 | 1.03 | 1.03 | $ 91,445.34 | 0.0004 | 0.04% |
| Provider 42 | 1.00 | 1.04 | 1.04 | $103,314.88 | -0.0011 | -0.11% |
| Provider 43 | 1.36 | 1.04 | 1.04 | $ 74,269.13 | 0.0026 | 0.25% |
| Provider 44 | 1.17 | 1.04 | 1.04 | $ 80,904.86 | 0.0018 | 0.17% |
| Provider 45 | 0.96 | 1.05 | 1.04 | $100,032.86 | -0.0007 | -0.07% |
| Provider 46 | 1.21 | 1.05 | 1.05 | $101,489.30 | -0.0009 | -0.08% |
| Provider 47 | 1.08 | 1.06 | 1.06 | $104,618.70 | -0.0013 | -0.12% |
| Provider 48 | 1.09 | 1.06 | 1.06 | $ 83,775.49 | 0.0014 | 0.13% |
| Provider 49 | 1.45 | 1.10 | 1.10 | $ 92,688.65 | 0.0003 | 0.02% |
| Provider 50 | 1.02 | 1.13 | 1.13 | $119,501.39 | -0.0032 | -0.28% |
| Provider 51 | 1.13 | 1.14 | 1.13 | $130,422.56 | -0.0046 | -0.41% |
| Provider 52 | 1.06 | 1.14 | 1.14 | $110,400.00 | -0.0020 | -0.18% |
| Provider 53 | 1.14 | 1.15 | 1.15 | $122,711.45 | -0.0036 | -0.31% |
| Provider 54 | 1.05 | 1.16 | 1.15 | $114,180.16 | -0.0025 | -0.22% |
| Provider 55 | 2.17 | 1.39 | 1.39 | $108,363.80 | -0.0018 | -0.13% |

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps—do not just name a method; what statistical analysis was used*)

Correlations and regression analysis utilized in both validity and the socioeconomic testing papers as well as the results in the Society of Actuaries study indicate that the statistical model used to adjust cost variation is effective. Additionally, because the commercial population's use of the healthcare system is so significantly different from the Medicaid and Medicare populations, through the benefits covered, the predominant conditions treated, and the prices of the services rendered, segmentation is required.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.*

***If stratified, skip to 2b4.9***

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

The Total Cost of Care measure uses the Johns Hopkins Adjusted Clinical Grouper (ACG) which adjusts for variation in risk profile using age, gender, and diagnosis (clinical risk adjustment). The measure is also limited by insurance coverage to commercial only. An evaluation between commercial and Medicaid covered members was also conducted in the socioeconomic testing, highlighting the variation in total cost (results included in 2b4.9.).

The non-risk adjusted PMPM coefficient of 0.62 in the table below indicates a high correlation between total cost and risk score.

| | Correlation Coefficient | |
|---|---|---|
| Metric | ACG | Non-Risk Adj PMPMs |
| Non-Risk Adj PMPM | 0.62 | 1.00 |
| Non-Risk Adj TCRRVs | 0.88 | 0.78 |
| ACG Risk Adj TCI | 0.03 | 0.79 |
| ACG Risk Adj RUI | 0.14 | 0.45 |
| Price | -0.09 | 0.57 |

| | R-Sqaured | |
|---|---|---|
| Metric | ACG | Non-Risk Adj PMPMs |
| Non-Risk Adj PMPM | 0.38 | 1.00 |
| Non-Risk Adj TCRRVs | 0.77 | 0.60 |
| ACG Risk Adj TCI | 0.00 | 0.62 |
| ACG Risk Adj RUI | 0.02 | 0.20 |
| Price | 0.01 | 0.33 |

Validity Paper (see page 5):
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188104.pdf

Socioeconomic Testing Paper (see page 4):
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

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

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

**2b4.9. Results of Risk Stratification Analysis**:

Detailed results can be found on page 4 and 5 of the socioeconomic testing paper:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

Using Census Tract Data the stratification results are shown in the far right column.

| Model | 1% Income Increase | 1% ACG increase | Commercial vs. Medicaid Membership |
|---|---|---|---|
| Total Reimbursement | $ (0.13) | $ 4.22 | $ 133.28 |
| Resource Use | $ 0.16 | $ 4.34 | $ (75.24) |
| Price | $ (0.28) | $ 0.07 | $ 205.36 |

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

Detailed results can be found on page 4 and 5 of the socioeconomic testing paper:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188119.pdf

Product contributed significantly with there being a $133 dollar difference in cost between commercial and Medicaid. The variation in resource use was much less, however, still significant with Medicaid covered members utilizing $75 more dollars of resources. The fact that Medicaid's cost per service is approximately half that of commercial rates drives the differences between the TCOC and Resource Use results.

**2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

________________________

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

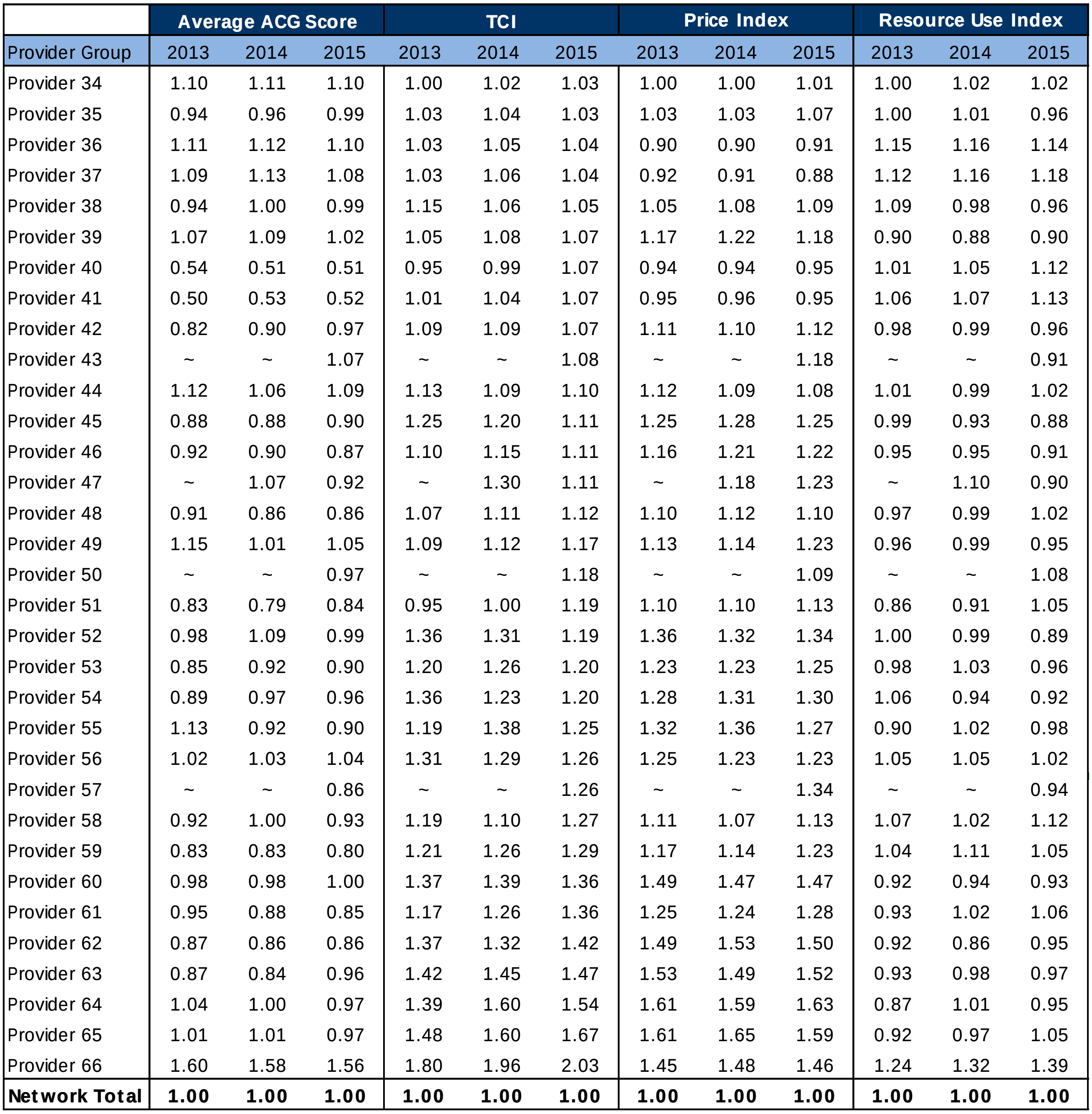
**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Performance is measured on an Index basis relative to 1.00 where each one point (0.01) variation from 1.00 (average) represents a 1% deviation from average. Statistical significance ranges of performance are not necessary as the measure is based on the full population. The results can be analyzed by percentile, percent from mean, standard deviation and clustering methods, this is dependent upon the business application of the measure.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

| | Average ACG Score | | | TCI | | | Price Index | | | Resource Use Index | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Provider Group | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 |
| Provider 01 | 1.11 | 1.09 | 1.09 | 0.87 | 0.84 | 0.84 | 0.89 | 0.91 | 0.89 | 0.98 | 0.93 | 0.93 |
| Provider 02 | ~ | 1.18 | 1.04 | ~ | 0.88 | 0.84 | ~ | 0.90 | 0.88 | ~ | 0.98 | 0.96 |
| Provider 03 | 0.85 | 0.86 | 0.88 | 0.93 | 0.86 | 0.85 | 0.95 | 0.94 | 0.93 | 0.98 | 0.91 | 0.92 |
| Provider 04 | 0.86 | 0.91 | 0.89 | 0.82 | 0.88 | 0.87 | 0.89 | 0.89 | 0.90 | 0.92 | 0.98 | 0.97 |
| Provider 05 | 1.27 | 1.15 | 1.12 | 0.93 | 0.86 | 0.87 | 1.14 | 1.11 | 1.06 | 0.81 | 0.78 | 0.82 |
| Provider 06 | 1.04 | 1.04 | 1.01 | 0.82 | 0.90 | 0.88 | 0.88 | 0.89 | 0.92 | 0.93 | 1.01 | 0.96 |
| Provider 07 | 1.08 | 1.08 | 1.05 | 0.92 | 0.92 | 0.89 | 0.95 | 0.92 | 0.92 | 0.98 | 1.00 | 0.97 |
| Provider 08 | 1.03 | 0.99 | 1.03 | 0.93 | 0.96 | 0.90 | 0.91 | 0.94 | 0.96 | 1.03 | 1.02 | 0.94 |
| Provider 09 | 1.00 | 1.04 | 1.06 | 0.86 | 0.86 | 0.90 | 0.88 | 0.89 | 0.91 | 0.98 | 0.97 | 1.00 |
| Provider 10 | 1.16 | 1.17 | 1.19 | 0.80 | 0.87 | 0.90 | 0.86 | 0.91 | 0.92 | 0.93 | 0.96 | 0.98 |
| Provider 11 | 1.19 | 1.35 | 1.42 | 1.02 | 0.93 | 0.91 | 1.01 | 0.96 | 0.91 | 1.00 | 0.97 | 1.00 |
| Provider 12 | 1.07 | 1.05 | 1.06 | 0.90 | 0.91 | 0.91 | 0.92 | 0.92 | 0.93 | 0.98 | 0.98 | 0.98 |
| Provider 13 | 1.01 | 1.06 | 1.06 | 0.95 | 0.95 | 0.92 | 0.91 | 0.93 | 0.93 | 1.05 | 1.02 | 0.99 |
| Provider 14 | 1.17 | 1.15 | 1.13 | 0.84 | 0.88 | 0.92 | 0.88 | 0.93 | 0.97 | 0.96 | 0.95 | 0.95 |
| Provider 15 | 0.91 | 0.94 | 0.95 | 0.84 | 0.97 | 0.92 | 0.94 | 0.98 | 1.00 | 0.89 | 0.99 | 0.92 |
| Provider 16 | 1.17 | 1.09 | 1.09 | 0.91 | 0.94 | 0.93 | 0.90 | 0.90 | 0.90 | 1.01 | 1.04 | 1.03 |
| Provider 17 | 1.14 | 1.14 | 1.14 | 0.89 | 0.85 | 0.93 | 0.89 | 0.87 | 0.95 | 1.00 | 0.98 | 0.97 |
| Provider 18 | 0.98 | 1.05 | 0.99 | 1.01 | 0.98 | 0.94 | 1.03 | 1.04 | 1.06 | 0.98 | 0.94 | 0.89 |
| Provider 19 | 0.90 | 0.88 | 0.86 | 0.95 | 0.92 | 0.94 | 0.94 | 0.93 | 0.94 | 1.01 | 0.98 | 1.01 |
| Provider 20 | 0.99 | 1.02 | 1.04 | 1.00 | 0.96 | 0.96 | 0.97 | 0.97 | 0.98 | 1.03 | 0.99 | 0.98 |
| Provider 21 | 0.82 | 0.84 | 0.85 | 0.98 | 1.00 | 0.96 | 0.95 | 0.94 | 0.95 | 1.04 | 1.07 | 1.01 |
| Provider 22 | 1.04 | 0.93 | 0.94 | 1.07 | 1.03 | 0.96 | 1.10 | 1.15 | 1.11 | 0.97 | 0.90 | 0.87 |
| Provider 23 | 0.88 | 0.96 | 0.94 | 1.09 | 0.98 | 0.96 | 0.96 | 0.97 | 0.95 | 1.14 | 1.01 | 1.01 |
| Provider 24 | 0.91 | 0.94 | 0.94 | 0.96 | 0.96 | 0.97 | 0.93 | 0.94 | 0.94 | 1.02 | 1.02 | 1.03 |
| Provider 25 | 1.20 | 1.12 | 1.20 | 0.94 | 0.84 | 0.97 | 0.94 | 0.93 | 0.99 | 1.00 | 0.90 | 0.99 |
| Provider 26 | 1.07 | 1.07 | 1.07 | 0.95 | 0.96 | 0.98 | 0.97 | 0.97 | 0.98 | 0.98 | 0.99 | 1.00 |
| Provider 27 | 1.06 | 1.04 | 1.03 | 1.02 | 0.96 | 0.98 | 0.91 | 0.88 | 0.87 | 1.12 | 1.09 | 1.12 |
| Provider 28 | 1.02 | 1.03 | 1.04 | 0.96 | 0.97 | 0.98 | 0.93 | 0.93 | 0.94 | 1.03 | 1.04 | 1.05 |
| Provider 29 | 0.96 | 1.02 | 1.03 | 1.12 | 1.07 | 1.01 | 1.10 | 1.08 | 1.08 | 1.02 | 0.99 | 0.94 |
| Provider 30 | 0.89 | 0.93 | 0.90 | 1.03 | 0.99 | 1.01 | 0.97 | 0.98 | 0.99 | 1.07 | 1.01 | 1.02 |
| Provider 31 | 1.01 | 0.98 | 0.98 | 1.03 | 1.01 | 1.01 | 1.06 | 1.10 | 1.12 | 0.97 | 0.91 | 0.91 |
| Provider 32 | 1.00 | 0.96 | 1.06 | 1.04 | 0.97 | 1.02 | 0.95 | 0.94 | 0.98 | 1.09 | 1.03 | 1.04 |
| Provider 33 | ~ | 0.95 | 1.11 | ~ | 1.00 | 1.02 | ~ | 0.88 | 0.89 | ~ | 1.14 | 1.15 |

The red line divides providers between above and below the average total cost index (1.00).



| | Average ACG Score | | | TCI | | | Price Index | | | Resource Use Index | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Provider Group | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 |
| Provider 34 | 1.10 | 1.11 | 1.10 | 1.00 | 1.02 | 1.03 | 1.00 | 1.00 | 1.01 | 1.00 | 1.02 | 1.02 |
| Provider 35 | 0.94 | 0.96 | 0.99 | 1.03 | 1.04 | 1.03 | 1.03 | 1.03 | 1.07 | 1.00 | 1.01 | 0.96 |
| Provider 36 | 1.11 | 1.12 | 1.10 | 1.03 | 1.05 | 1.04 | 0.90 | 0.90 | 0.91 | 1.15 | 1.16 | 1.14 |
| Provider 37 | 1.09 | 1.13 | 1.08 | 1.03 | 1.06 | 1.04 | 0.92 | 0.91 | 0.88 | 1.12 | 1.16 | 1.18 |
| Provider 38 | 0.94 | 1.00 | 0.99 | 1.15 | 1.06 | 1.05 | 1.05 | 1.08 | 1.09 | 1.09 | 0.98 | 0.96 |
| Provider 39 | 1.07 | 1.09 | 1.02 | 1.05 | 1.08 | 1.07 | 1.17 | 1.22 | 1.18 | 0.90 | 0.88 | 0.90 |
| Provider 40 | 0.54 | 0.51 | 0.51 | 0.95 | 0.99 | 1.07 | 0.94 | 0.94 | 0.95 | 1.01 | 1.05 | 1.12 |
| Provider 41 | 0.50 | 0.53 | 0.52 | 1.01 | 1.04 | 1.07 | 0.95 | 0.96 | 0.95 | 1.06 | 1.07 | 1.13 |
| Provider 42 | 0.82 | 0.90 | 0.97 | 1.09 | 1.09 | 1.07 | 1.11 | 1.10 | 1.12 | 0.98 | 0.99 | 0.96 |
| Provider 43 | ~ | ~ | 1.07 | ~ | ~ | 1.08 | ~ | ~ | 1.18 | ~ | ~ | 0.91 |
| Provider 44 | 1.12 | 1.06 | 1.09 | 1.13 | 1.09 | 1.10 | 1.12 | 1.09 | 1.08 | 1.01 | 0.99 | 1.02 |
| Provider 45 | 0.88 | 0.88 | 0.90 | 1.25 | 1.20 | 1.11 | 1.25 | 1.28 | 1.25 | 0.99 | 0.93 | 0.88 |
| Provider 46 | 0.92 | 0.90 | 0.87 | 1.10 | 1.15 | 1.11 | 1.16 | 1.21 | 1.22 | 0.95 | 0.95 | 0.91 |
| Provider 47 | ~ | 1.07 | 0.92 | ~ | 1.30 | 1.11 | ~ | 1.18 | 1.23 | ~ | 1.10 | 0.90 |
| Provider 48 | 0.91 | 0.86 | 0.86 | 1.07 | 1.11 | 1.12 | 1.10 | 1.12 | 1.10 | 0.97 | 0.99 | 1.02 |
| Provider 49 | 1.15 | 1.01 | 1.05 | 1.09 | 1.12 | 1.17 | 1.13 | 1.14 | 1.23 | 0.96 | 0.99 | 0.95 |
| Provider 50 | ~ | ~ | 0.97 | ~ | ~ | 1.18 | ~ | ~ | 1.09 | ~ | ~ | 1.08 |
| Provider 51 | 0.83 | 0.79 | 0.84 | 0.95 | 1.00 | 1.19 | 1.10 | 1.10 | 1.13 | 0.86 | 0.91 | 1.05 |
| Provider 52 | 0.98 | 1.09 | 0.99 | 1.36 | 1.31 | 1.19 | 1.36 | 1.32 | 1.34 | 1.00 | 0.99 | 0.89 |
| Provider 53 | 0.85 | 0.92 | 0.90 | 1.20 | 1.26 | 1.20 | 1.23 | 1.23 | 1.25 | 0.98 | 1.03 | 0.96 |
| Provider 54 | 0.89 | 0.97 | 0.96 | 1.36 | 1.23 | 1.20 | 1.28 | 1.31 | 1.30 | 1.06 | 0.94 | 0.92 |
| Provider 55 | 1.13 | 0.92 | 0.90 | 1.19 | 1.38 | 1.25 | 1.32 | 1.36 | 1.27 | 0.90 | 1.02 | 0.98 |
| Provider 56 | 1.02 | 1.03 | 1.04 | 1.31 | 1.29 | 1.26 | 1.25 | 1.23 | 1.23 | 1.05 | 1.05 | 1.02 |
| Provider 57 | ~ | ~ | 0.86 | ~ | ~ | 1.26 | ~ | ~ | 1.34 | ~ | ~ | 0.94 |
| Provider 58 | 0.92 | 1.00 | 0.93 | 1.19 | 1.10 | 1.27 | 1.11 | 1.07 | 1.13 | 1.07 | 1.02 | 1.12 |
| Provider 59 | 0.83 | 0.83 | 0.80 | 1.21 | 1.26 | 1.29 | 1.17 | 1.14 | 1.23 | 1.04 | 1.11 | 1.05 |
| Provider 60 | 0.98 | 0.98 | 1.00 | 1.37 | 1.39 | 1.36 | 1.49 | 1.47 | 1.47 | 0.92 | 0.94 | 0.93 |
| Provider 61 | 0.95 | 0.88 | 0.85 | 1.17 | 1.26 | 1.36 | 1.25 | 1.24 | 1.28 | 0.93 | 1.02 | 1.06 |
| Provider 62 | 0.87 | 0.86 | 0.86 | 1.37 | 1.32 | 1.42 | 1.49 | 1.53 | 1.50 | 0.92 | 0.86 | 0.95 |
| Provider 63 | 0.87 | 0.84 | 0.96 | 1.42 | 1.45 | 1.47 | 1.53 | 1.49 | 1.52 | 0.93 | 0.98 | 0.97 |
| Provider 64 | 1.04 | 1.00 | 0.97 | 1.39 | 1.60 | 1.54 | 1.61 | 1.59 | 1.63 | 0.87 | 1.01 | 0.95 |
| Provider 65 | 1.01 | 1.01 | 0.97 | 1.48 | 1.60 | 1.67 | 1.61 | 1.65 | 1.59 | 0.92 | 0.97 | 1.05 |
| Provider 66 | 1.60 | 1.58 | 1.56 | 1.80 | 1.96 | 2.03 | 1.45 | 1.48 | 1.46 | 1.24 | 1.32 | 1.39 |
| **Network Total** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

The Total Cost of Care measure can effectively identify variation in performance levels.

Practically meaningful difference in performance will vary by use of the measures. This is because some uses may have a higher threshold for differences. For example, a 10% difference in performance when the result is used for public reporting could be very meaningful in terms of provider patient growth and retention strategies. The same 10% difference may not be as meaningful when using the measures internally for improvement work and identification of a work plan.

The following will give a general sense of the dispersion of the scoring:

Out of the 66 provider groups measured in Total Cost of Care:

- 28 were better than average
- 10 were 10% better than average
- 23 were 10% higher than average
- 33 were within 10% of the average

______________________________________

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**
***If only one set of specifications, this section can be skipped*.**

__Note__*: This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps—do not just name a method; what statistical analysis was used*)

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

**2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

______________________________________

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

This is a full population-based measure, all data is included in the measure.

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; __if no empirical sensitivity analysis__, identify the approaches for handling missing data that were considered and pros and cons of each*)

This is a full population-based measure, all data is included in the measure.

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

This is a full population-based measure, all data is included in the measure.

## Feasibility

F.1. **Byproduct of Care Processes**

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

F.1.1. **Data Elements Generated as Byproduct of Care Processes.**

Other

If other: Health Plan Claims data system

F.2. **Electronic Sources**

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

F.2.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*)

ALL data elements are in defined fields in a combination of electronic sources

F.2.1a. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.

F.2.2. **If this is an eMeasure,** provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.

**Attachment:**

F.3. **Data Collection Strategy**

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

F.3.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

Since endorsement we have received some general feedback regarding implementation of the measure. This has helped shape some of the materials and additional testing we've conducted since the measures were first released. HealthPartners has organized a public-facing website with several resources and technical documentation, including toolkits for external organizations to download necessary tools to run the measure, free of charge. In addition, HealthPartners uses SAS to run the measure and not every organization has or uses this software. To address this, HealthPartners organized non-SAS user instructions. By creating these resources and software and putting them in the public domain it has resulted in expanded use. A few users have successfully implemented the NQF-endorsed Total Cost of Care measure according to the specifications, however they are using their previously purchased risk grouper (not ACG).

F.3.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set,**

**risk model, programming code, and algorithm)?**

The measure and software are available free of charge at www.healthpartners.com/tcoc;
The TCOC measure download options are available at: https://www.healthpartners.com/hp/about/tcoc/toolkit/index.html
The ACG System is widely available within the public and private sectors in the US and abroad.(Bibliography: http://acg.jhsph.org/index.php/resource-center-83/acg-bibliography) The pricing for the ACG System varies for commercial and government entities but is generally based on a per member per year license that is tiered to provide lower per member costs for larger entities. For a commercial plan there is a base fee of $27,000 annually with incremental costs between $0.04 and $0.40 per member per year based on volume, which is inclusive of both license fees and support. Discounted fees are available for government entities and other grant funded not-for-profit entities. Additionally, Johns Hopkins offers research licenses for a very modest cost for academic users incorporating ACGs into published research: http://www.acg.jhsph.org/index.php?option=com_content&view=article&id=137&Itemid=94

The ACG System technical guide is available here:
https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/cntrb_035024.pdf

F.3.3. **If there are any fees associated with the use of this measure as specified, attach the fee schedule here. (Save file as: F3_3_FeeSchedule)**
FeeScheduleTemplate_Proprietary_Fees_V2.0SubmissionForm-Johns_Hopkins_ACG_System_2016-11.xlsx

## Usability and Use

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

*NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.*

U.1.1. **Current and Planned Use**

| Specific Plan for Use | Current Use (for current use provide URL) |
|---|---|
| | Public Reporting<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf<br><br>Public Health/Disease Surveillance<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf<br><br>Payment Program<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf<br><br>Quality Improvement (external benchmarking to organizations)<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf<br><br>Quality Improvement (Internal to the specific organization)<br>See URL<br>https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf |

U.1.2. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Since endorsement in 2012, uptake of the Total Cost of Care measure has expanded across 37 states in the country and used by both national and regional organizations (Coverage). The measure has been used in accountability applications and publicly reported in multiple states for driving improvement.

The following link highlights organizations across the country that have adopted the measure and are currently using it for at least one of the uses noted above, including some crossover of multiple uses for some organizations. The URLs of the specific programs are included within the link below to appropriately capture each organization's purpose described in their own words.

Because some of the organizations are using the measure for multiple uses, we have included them based on their predominant category.

https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf

Public Reporting

HealthPartners – Public Reporting, Payment Program, Quality Improvement with Benchmarking

- As a health plan, HealthPartners uses the Total Cost of Care measure to incentivize providers to meet Triple Aim goals, optimizing health and patient experience while improving affordability. HealthPartners publicly reports provider group cost results for purposes of transparency for employers, providers, and consumers. The cost results are paired with Total Resource Use, quality and experience metrics to promote quality improvement with benchmarking across providers.

- HealthPartners has shared savings payment agreements with over 85% of its primary care providers which has increased provider engagement and sharing of appropriate risk as a partnership to lower cost for providers and patients while maintaining quality and experience. Additionally, in conjunction with the Total Resource Use measure, HealthPartners has begun building upon it by implementing new payment reform models that align incentives among specialists and hospitals to support shared savings with primary care. The new methods include bundled payments for episodes of care as well as models that move away from fee for service and promote coordination of care.

MN Community Measurement – Public Reporting, Quality Improvement with Benchmarking

- Beginning in 2014, MNCM was the first community collaborative in the nation to publicly report Total Cost of Care data by provider group in Minnesota using HealthPartners endorsed Total Cost of Care measure. Through their multi-stakeholder collaborative process they were able to collect cost data from four health plans and publicly spread the use of the measure to all provider groups in Minnesota, promoting transparency.

Network for Regional Healthcare Improvement – Public Reporting, Quality Improvement with Benchmarking, Quality Improvement

- Eleven regions are part of a project to develop a standardized method of reporting total cost and resource use by using the HealthPartners endorsed measures. During 2015, seven regions were able to share healthcare cost information on over 5 million patients attributed to 20,000 individual physicians through practice level reports. Their work is described in detail in the provided link.

Payment Program

The Alliance

- Utilizes the measures for provider contracting and incentives.

Public Health/Disease Surveillance

The University of Iowa

- Research evaluation for assessing state health system transformation under the State Innovation Model initiative.

Quality Improvement with Benchmarking

Maine Health Management Coalition – regional collaborative

- Commercial premium costs will be measured against benchmarks using the TCOC and Resource Use measures with plans for future public reporting.

Oregon Health Care Quality Corporation – regional collaborative
- In 2015, Q Corp released Clinic Comparison Reports featuring cost, utilization and quality measures to over 150 primary care clinics in Oregon.

HealthInsight and Utah Department of Health, Washington Health Alliance – regional collaboratives
- Regional collaboratives participating in the Network for Regional Healthcare Improvement's project to develop a standardized method of reporting total cost and resource use.

Center for Improving Value in Health Care (CIVHC) – regional collaborative
- Recently began providing results to Colorado primary care groups to help them see how their practice patterns compare.

Midwest Health Initiative – regional collaborative
- Shares data with physician groups and practice sites through community reports with future plans for public reporting.

Quality Improvement

Provider Groups in Minnesota
- Having payment agreements with HealthPartners, several provider groups see the value and are actively engaged in utilizing the Total Cost of Care measure. They shared with us how they are using the measure within their own practice and their letters of support are included in the link.

American Hospital Association
- Partnering with HealthPartners to develop a pilot of the measure across their constituents for broader use.

Priority Health
- Evaluating practice efficiencies and pricing fluctuations across Accountable Care Networks.

Providence Health Plans
- Provide efficiency profiling and increasing engagement for improvement and better referral decision making.

Onpoint Health Data
- State organization are utilizing the data for program evaluation.

National Quality Measures Clearinghouse (NQMC) from Agency for Healthcare Research and Quality (AHRQ) reported the following usage between 3/1/15 – 2/29/16
- 5,815 page views for the Total Cost of Care Measure

U.1.3. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not applicable

U.1.4. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not applicable

U.2.1. **Progress on Improvement. (Not required for initial endorsement unless available.) Performance results on this measure (current and over time) should be provided in IM.2.2 and IM.2.4.**
**Discuss:**

- **Purpose Progress (trends in performance results)**
- **Geographic area and number and percentage of accountable entities and patients included**

https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/entry_188106.rtf

HealthPartners Medical Group, Park Nicollet Health Services, Essentia Health, CentraCare Health and Fairview are provider groups in Minnesota who are highlighted as engaged users of the measure and who have seen improvement in their care practices. The details they've shared and the strategies they've implemented to lower cost are included in the link provided.

Since endorsement in 2012, there has been a large increase in the number of users who have adopted the Total Cost of Care measure, in conjunction with the Total Resource Use measure, resulting in improvement through greater transparency. An increase in transparency brings an awareness to the rising healthcare costs in our communities and has helped users pinpoint areas for improvement and define strategies to reduce those costs.

HealthPartners has also organized a public-facing website with several resources and technical documentation, including toolkits for external organizations to download the necessary tools to run the measure, free of charge. In addition, HealthPartners has created instructions and toolkits for both SAS and non-SAS users. By creating these resources and software and putting them in the public domain it has resulted in expanded use.

The following link includes details of one specific example demonstrating improvement and features the Northwest Metro Alliance which serves more than 300,000 people receiving care at 9 different clinics and one hospital along with its affiliated specialists. The Alliance's medical cost increases were more than 31 percent lower than the Twin Cities metro average for Commercial patients since they adopted the Total Cost of Care and Resource Use measures in 2010.

Link to and post on website:

https://www.allinahealth.org/uploadedFiles/Content/Customer_Service/Billing_and_insurance/Northwest-Metro-Alliance-5-year-results.pdf

U.2.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not applicable

U.3.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
HealthPartners mitigates risk through the following steps:

- Claims data integrity procedures prior to loading data warehouse through HealthPartners Data Integrity Dept.
- Internal Audit Dept. review of processes & procedures for generating measure
- Provider contracts allow ability to request external audit
- HealthPartners Provider Measurement Policy allows for a 45-day comment period before results are used in any business applications (incentive, public display, etc). Any identified errors ore issues are resolved & corrected

To our knowledge we are not aware of negative unintended consequences from other organizations utilizing the measure.

Since endorsement we have received some general feedback regarding implementation of the measure. This has helped shape some of the materials and additional testing we've conducted since the measure was first released. HealthPartners has organized a public-facing website with several resources and technical documentation, including toolkits for external organizations to download necessary tools to run the measure, free of charge. In addition, HealthPartners uses SAS to run the measure and not every organization has or uses this software. To address this, HealthPartners organized non-SAS user instructions. By creating these resources and software and putting them in the public domain it has resulted in expanded use. A couple of external users have shared feedback on possible barriers with funding of the ACG grouper when the organization has already invested in a different grouper.

## Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

H.1. **Relation to Other NQF-endorsed Measures**
If there are related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

H.1.1. **List of related or competing measures (selected from NQF-endorsed measures)**

H.1.2. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
Not applicable

**H.2. Harmonization**

H.2.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
**Are the measure specifications completely harmonized?**

H.2.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

H.3. **Competing Measure(s)**

H.3.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
**Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)**
Not applicable

## Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** HealthPartners
Co.2 **Point of Contact:** Sue, Knudson, Susan.M.Knudson@healthpartners.com, 952-883-6185-
Co.3 **Measure Developer if different from Measure Steward:** HealthPartners
Co.4 **Point of Contact:** Sue, Knudson, Susan.M.Knudson@healthpartners.com, 952-883-6185-

## Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
List the workgroup/panel members' names and organizations.
Describe the members' role in measure development.

**Measure Developer/Steward Updates and Ongoing Maintenance**
Ad.2 **Year the measure was first released:** 2003
Ad.3 **Month and Year of most recent revision:** 06, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 06, 2017

Ad.6 **Copyright statement:** © 2016 HealthPartners. Reprints allowed for noncommercial purposes only if this copyright notice is prominently included and HealthPartners is given clear attribution as the copyright owner.
Ad.7 **Disclaimers:** Total Cost of Care and Total Resource Use are licensed free of charge with supporting implementation tools at the following website:
www.healthpartners.com/tcoc

Ad.8 **Additional Information/Comments:** HealthPartners public Total Cost of Care and Total Resource Use site
www.healthpartners.com/tcoc

For purposes of the National Quality Forum Measure Maintenance Review for Endorsed HealthPartners Measures.

www.healthpartners.com/tcoc-documents

HealthPartners®

# Appendix A

Cost and Resource Use Project 2016-2017

National Quality Forum 2012 Measure Endorsement

Total Cost of Care (NQF#1604)
Total Resource Use (NQF#1598)

# 2012 Submission Technical Documentation: Reliability and Validity Testing

HealthPartners' Total Cost of Care and Total Resource Use measures use the same measurement criteria except for the costing method and are considered complementary to each other.

Appendix A supports the Measure Testing Attachments for Total Cost of Care and Total Resource Use Measure Maintenance. The methodology used for both submissions is consistent. Results from the prior testing period using earlier dates of services are included on the following pages.

Results from both testing periods indicate the Total Cost of Care and Total Resource Use measures are both reliable and valid.

HealthPartners®

HealthPartners Technical Documentation

# Total Cost of Care
# Bootstrap Reliability Analysis

## Purpose

Determine the reliability of the Total Cost of Care (TCI) measure.

## Table of Contents

## *Overview of Analysis*

Total Cost of Care (TCI) is a measure of a provider's effectiveness of managing their primary care attributed population across the care continuum. The TCI measure was applied to HealthPartners' primary care metro providers as per the measure specifications and results were calculated for 2007, 2008, and 2009.

The reliability testing demonstrates the repeatability of producing the same results a high proportion of the time. To measure the reliability of the TCI measure a 90% random sample and a bootstrapping technique were employed. In these methods, reliability is measured as the mean of the variance between sampling iterations and the actual results.

In addition, the TCI measure was analyzed over time to demonstrate stability and sensitivity to provider changes or improvement initiatives.

These methods were chosen as they represent the measure intent, which is that the TCI measure represents providers' average total cost of care across their population. Since the measure is aggregated to the provider group level there is no need to quantify the variability at the member level into the evaluation.

In the 90% random sample method, the members that were attributed to a provider group were randomly sampled at the 90% membership level without replacement. This technique was employed to simulate variation within a provider group by leveraging their own population and case-mix. This method gives an indication as to the repeatability of the measure by comparing how closely the actual total cost measure is to the 90% sampled averages and simulates any potential member selection bias.

In the bootstrapping method members that were attributed to a provider group were randomly selected with replacement. This method maximizes variation around a provider group's total cost of care as each randomly selected iteration (sample populations) does not truly represent the provider's case mix of patients. This method was performed in the same fashion as above to support and validate the results found in the 90% sample method.

## *Overall Conclusions*

- The differences between provider Actual TCI results and both the 90% sample and bootstrap mean results are very small.
    - Ranging from -0.0069 to 0.00083 in the 90% sample in 2009.
    - Ranging from -0.00067 to 0.00252 in the bootstrap in 2009.
    - These results indicate that the TCIs for each provider group are repeatable and consistent.
- A provider's performance is relatively consistent across all three years with an average difference of 0.031.
    - These differences in provider performance over time occur because of known changes in fee schedules, collaborating provider usage and resource use saving initiatives can account for the differences.
    - Since the measure is designed to capture and reflect changes in these areas, we expect to see some explainable variability within a provider group over time.

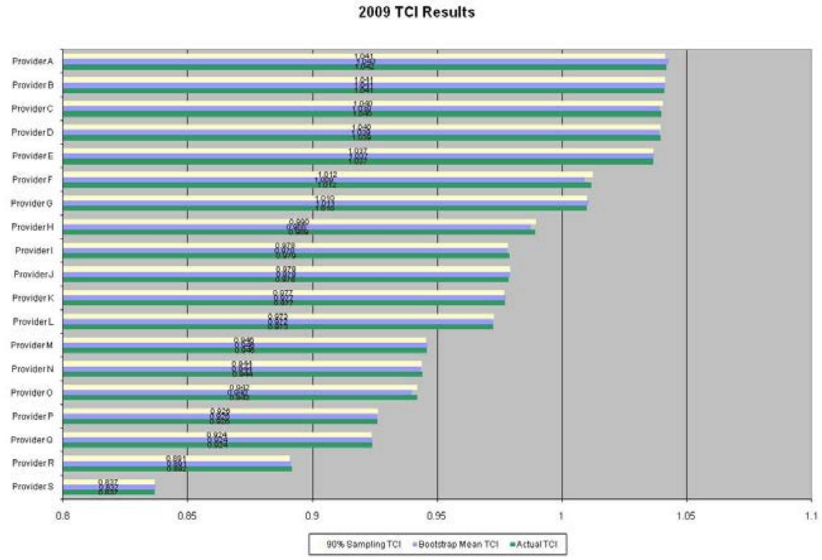
## *Methodology*

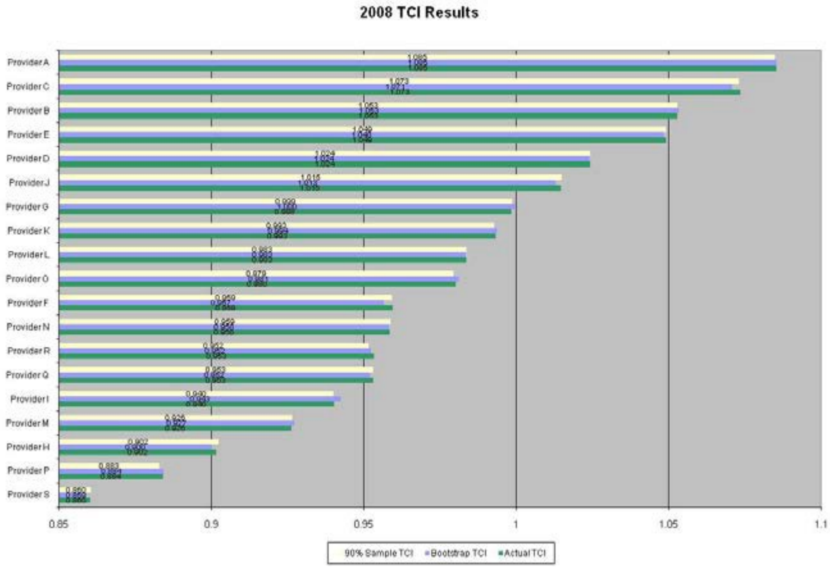
In the 90% sample method, 90% of attributed provider group members were randomly selected, without replacement. A 90% sample was used despite having the full health plan provider population, to simulate any potential member selection bias. The sampling process was performed using the SAS PROC SURVEYSELECT procedure with the Simple Random Sample (SRS) option. This method allows for each attributed member to be selected only one time until 90% of the total provider population has been reached. The 90% sampling process was repeated 500 times for each provider group and year analyzed. Attributed members' total costs were aggregated in each sample to produce 500 TCI results for each provider group for each year (see Figure 1 in the definitions section for more information). Once the 500 samples were created for each provider group, the total costs of care of each sample for each provider group were compared to the metro average to produce risk adjusted indices. The Total Cost indices from each of the sampling iterations for each provider group/year were then compared to the actual TCI indices for each provider group/year and the mean variance was computed.

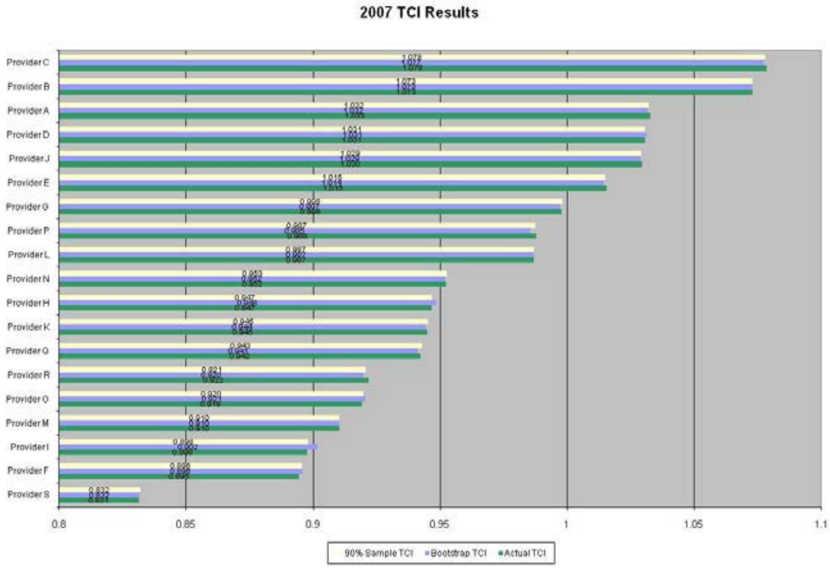
To perform the bootstrap, the SAS PROC SURVEYSELECT procedure with the Unrestricted Random Sample option for full replacement utilized to create a series of random samples for each provider group being measured. Full replacement means that one observation is drawn at random, recorded, and then placed back into the data pool so that it can be drawn again if randomly selected. The numbers of records sampled are drawn such that the samples created are the same size as the original number of attributed members for the provider group. In this way, it is theoretically possible (although virtually improbable) to produce a sample of size n that could consist of the same record drawn n times in a row. This was done to artificially maximize the variance within the defined populations. This sample process was performed 500 times for each year and provider group being analyzed, to produce 500 sets of risk adjusted Total Cost of Care results for each provider for each year (see Figure 2 in the definitions section for more information). The Total Cost indices from each of the sampling iterations for each provider group/year were then compared to the actual TCI indices for each provider group/year and the mean variance was computed.

## *Bootstrap and 90% Random Sample*

The mean TCI results from the bootstrap and 90% samples compared to the actual TCI results for each provider group and year are displayed in the tables and graphs on the following pages.





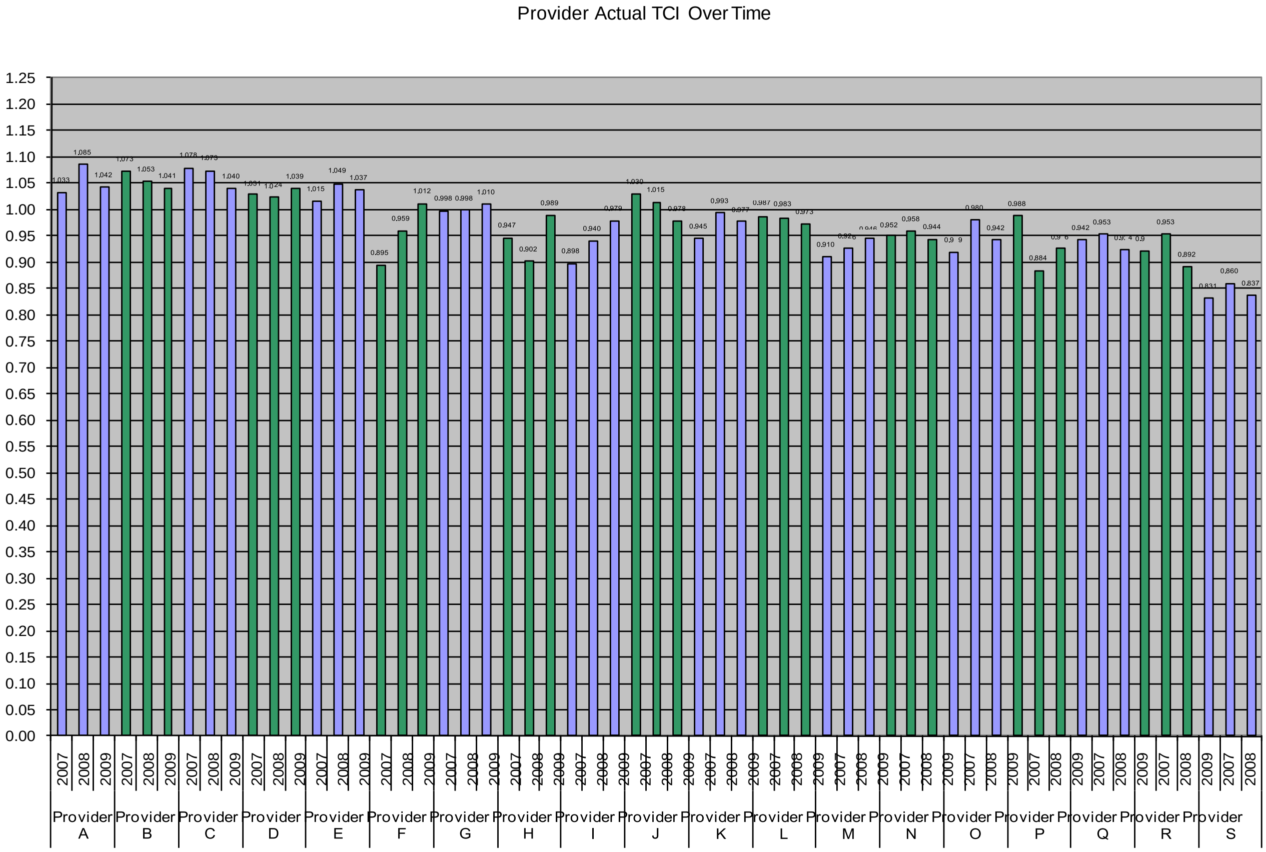


## *Bootstrap and 90% Random Sample Results*

- The differences between provider Actual TCI results and both the 90% sample and bootstrap mean results are very small ranging from -0.0069 to 0.00083 in the 90% sample to -0.00067 to 0.00252 in the bootstrap in 2009.
- The results indicate that the TCIs for each provider group are repeatable and consistent.

## TCI Consistency Over Time

The TCI results are displayed from 2007 through 2009 for the HealthPartners Primary Care Metro Network. The measure differentiates between providers however they remain relatively consistent over time. The factors that drive variation between years within a provider are cost per unit control and resource use management.


Provider Actual TCI Over Time

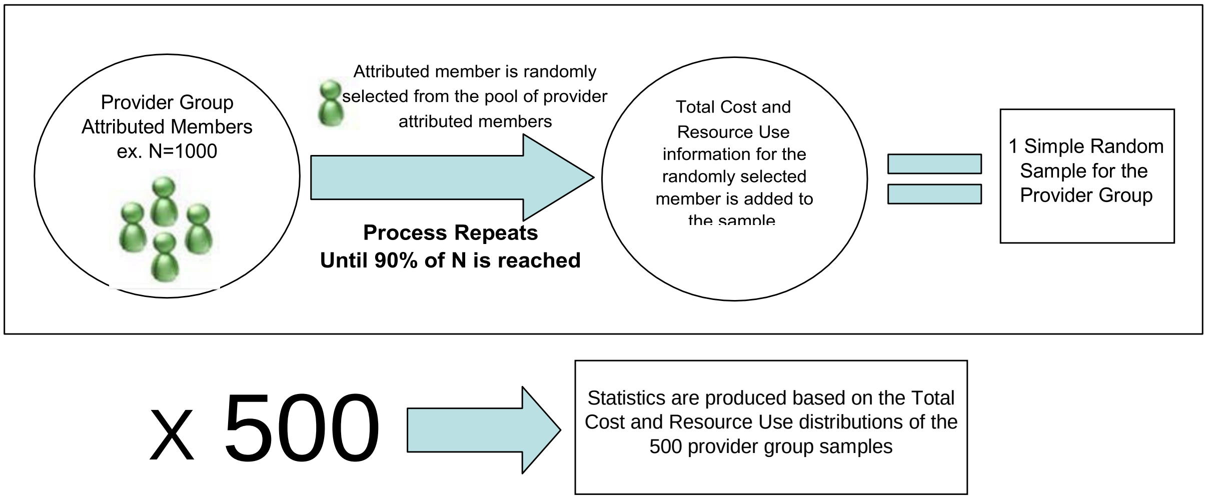
## TCI Consistency Over Time Results

A provider's relative performance is relatively consistent across all three years with an average difference of 0.031.

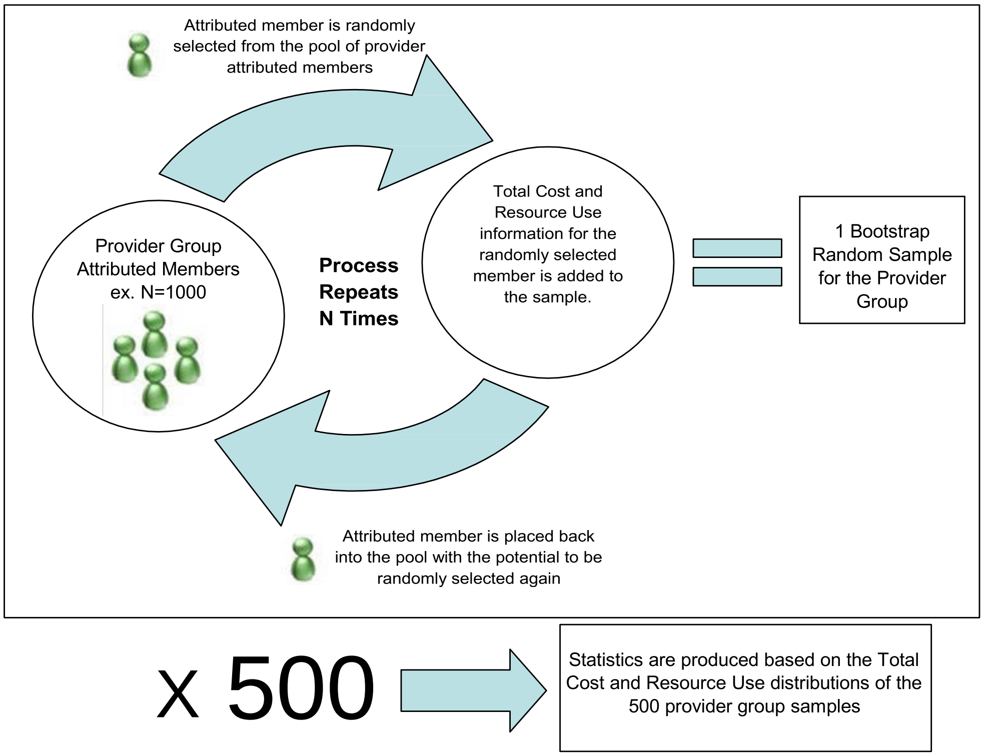
- These differences in provider performance over time occur because of known changes in fee schedules, collaborating provider usage and resource use saving initiatives can account for the differences.
- Since the measure is designed to capture and reflect changes in these areas, we expect to see some explainable variability within a provider group over time.

## *Definitions and Examples*

### Figure 1: 90% Sampling – Simple Random Sample Without Replacement


Provider Group Attributed Members ex. N=1000
Attributed member is randomly selected from the pool of provider attributed members
Process Repeats Until 90% of N is reached
Total Cost and Resource Use information for the randomly selected member is added to the sample
1 Simple Random Sample for the Provider Group
x 500
Statistics are produced based on the Total Cost and Resource Use distributions of the 500 provider group samples

### Figure 2: Bootstrap Sampling – Unrestricted Random Sampling With Full Replacement


Attributed member is randomly selected from the pool of provider attributed members
Provider Group Attributed Members ex. N=1000
Process Repeats N Times
Total Cost and Resource Use information for the randomly selected member is added to the sample.
1 Bootstrap Random Sample for the Provider Group
Attributed member is placed back into the pool with the potential to be randomly selected again
x 500
Statistics are produced based on the Total Cost and Resource Use distributions of the 500 provider group samples

Appendix A

HealthPartners®

HealthPartners Technical Documentation

# Total Resource Use Bootstrap Reliability Analysis

## Purpose

Determine the reliability of the Resource Use Index (RUI) measure.

## Table of Contents

## Overview of Analysis

Resource Use Index (RUI) is a measure of a provider's effectiveness of managing their primary care attributed population across the care continuum. The RUI measure was applied to HealthPartners primary care metro providers as per the measure specifications and results were calculated for 2007, 2008, and 2009.

The reliability testing demonstrates the repeatability of producing the same results a high proportion of the time. To measure the reliability of the RUI measure a 90% random sample and a bootstrapping technique were employed. In these methods, reliability is measured as the mean of the variance between sampling iterations and the actual results.

In addition, the RUI measure was analyzed over time to demonstrate stability and sensitivity to provider changes or improvement initiatives.

These methods were chosen as they represent the measure intent, which is that the RUI measure represents providers' average resource use across their population. Since the measure is aggregated to the provider group level there is no need to quantify the variability at the member level into the evaluation.

In the 90% random sample method, the members that were attributed to a provider group were randomly sampled at the 90% membership level without replacement. This technique was employed to simulate variation within a provider group by leveraging their own population and case-mix. This method gives an indication as to the repeatability of the measure by comparing how closely the actual resource use measure is to the 90% sampled average and simulates any potential member selection bias.

In the bootstrapping method members that were attributed to a provider group were randomly selected with replacement. This method maximizes variation around a provider group's resource use as each randomly selected iteration (sample populations) does not truly represent the provider's case mix of patients. This method was performed in the same fashion as above to support and validate the results found in the 90% sample method.

## Overall Conclusions

- The differences between provider Actual RUI results and both the 90% sample and bootstrap mean results are very small.
    - Ranging from -0.00449 to 0.00125 in the 90% sample in 2009.
    - Ranging from -0.00473 to 0.00105 in the bootstrap in 2009.
    - These results indicate that the RUIs for each provider group are repeatable and consistent.
- A provider's performance is relatively consistent across all three years with an average difference in RUI between 2008 and 2009 of 0.0125.
    - These differences in provider performance over time occur because of known changes in fee schedules, collaborating provider usage and resource use saving initiatives can account for the differences.
    - Since the measure is designed to capture and reflect changes in these areas, we expect to see some explainable variability within a provider group over time.

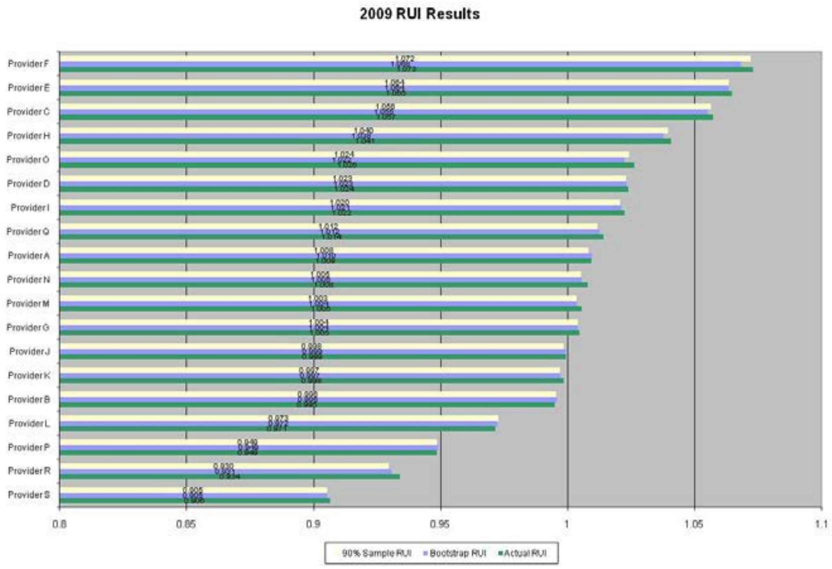
## Methodology

In the 90% sample method, 90% of attributed provider group members were randomly selected, without replacement. A 90% sample was used despite having the full health plan provider population, to simulate any potential member selection bias. The sampling process was performed using the SAS PROC SURVEYSELECT procedure with the Simple Random Sample (SRS) option. This method allows for each attributed member to be selected only one time until 90% of the total provider population has been reached. The 90% sampling process was repeated 500 times for each provider group and year analyzed. Attributed members' resource use was aggregated in each sample to produce 500 RUI results for each provider group for each year (see Figure 1 in the definitions section for more information). Once the 500 samples were created for each provider group, the resource use of each sample for each provider group was compared to the metro average to produce a risk adjusted index. The Resource Use Index from each of the sampling iterations for each provider group/year was then compared to the actual RUI for each provider group/year and the mean variance was computed.

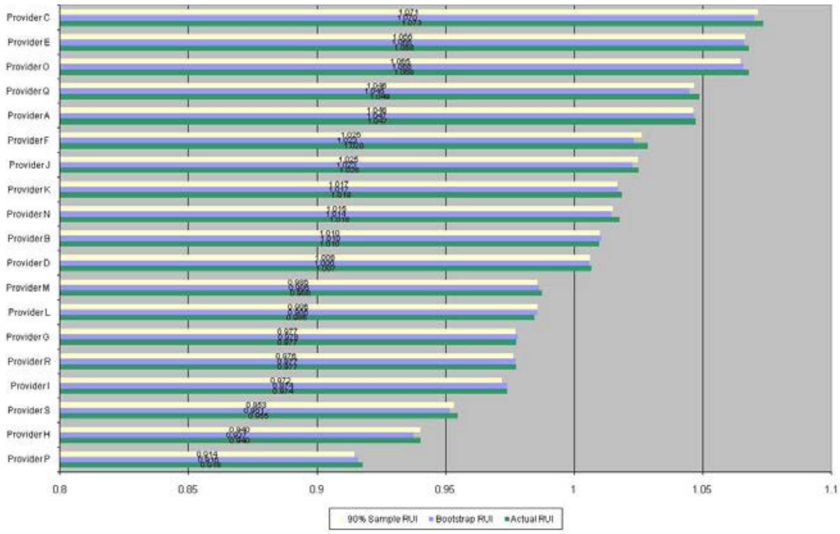
To perform the bootstrap, the SAS PROC SURVEYSELECT procedure with the Unrestricted Random Sample option for full replacement utilized to create a series of random samples for each provider group being measured. Full replacement means that one observation is drawn at random, recorded, and then placed back into the data pool so that it can be drawn again if randomly selected. The numbers of records sampled are drawn such that the samples created are the same size as the original number of attributed members for the provider group. In this way, it is theoretically possible (although virtually improbable) to produce a sample of size n that could consist of the same record drawn n times in a row. This was done to artificially maximize the variance within the defined populations. This sample process was performed 500 times for each year and provider group being analyzed, to produce 500 sets of risk adjusted Resource Use results for each provider for each year (see Figure 2 in the definitions section for more information). The Resource Use Index from each of the sampling iterations for each provider group/year was then compared to the actual RUI for each provider group/year and the mean variance was computed.

## Bootstrap and 90% Random Sample

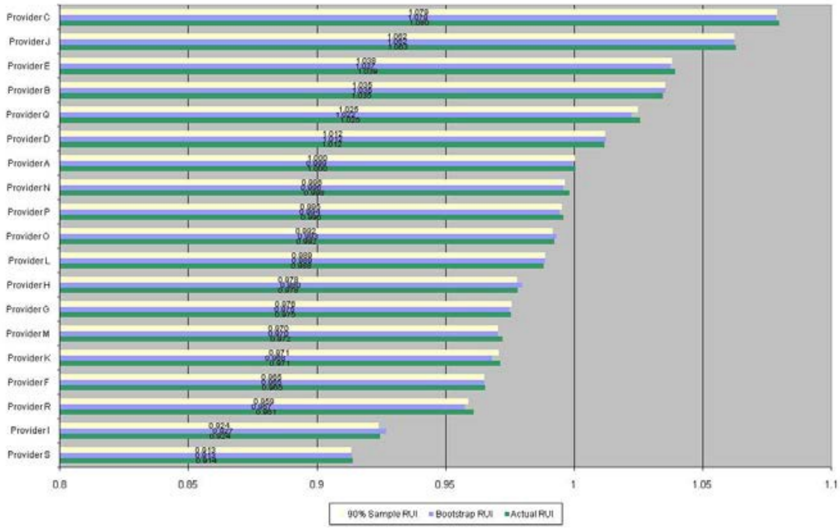
The mean Resource Use result from the bootstrap and 90% samples compared to the actual Resource Use result for each provider group and year is displayed in the tables and graphs on the following pages.



2008 RUI Results



2007 RUI Results



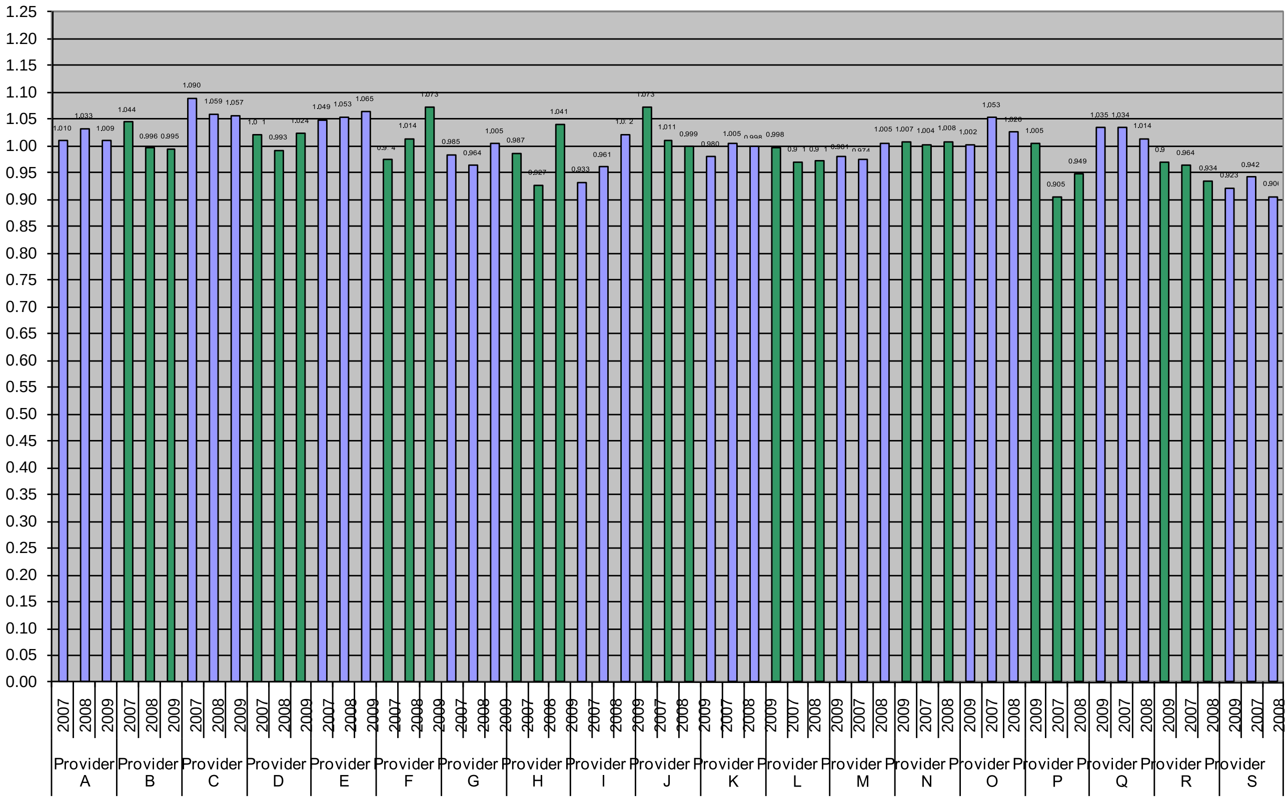
## *Bootstrap and 90% Random Sample Results*

- The differences between provider Actual RUI results and both the 90% sample and bootstrap mean results are very small ranging from -0.00449 to 0.00125 in the 90% sample to -0.00473 to 0.00105 in the bootstrap in 2009.
- The results indicate that the RUIs for each provider group are repeatable and consistent.

## RUI Consistency Over Time

The Resource Use results are displayed from 2007 through 2009 for the HealthPartners Primary Care Metro Network. The measure differentiates between providers however they remain relatively consistent over time. The factor that drives variation between years within a provider is resource use management.

Provider Actual RUI Over Time



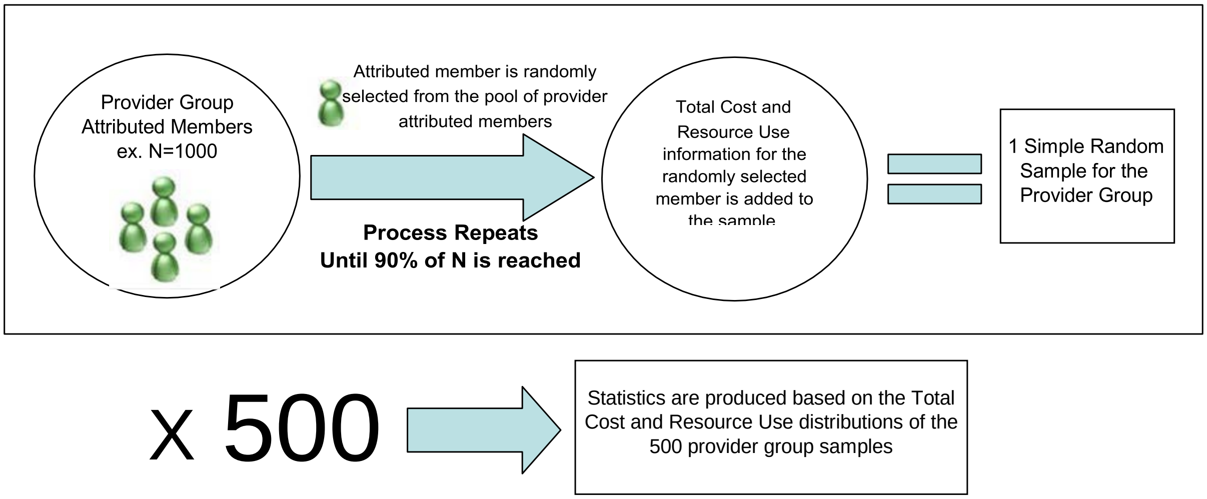
## RUI Consistency Over Time Results

A provider's relative performance is relatively consistent across all three years with an average difference of 0.0125.

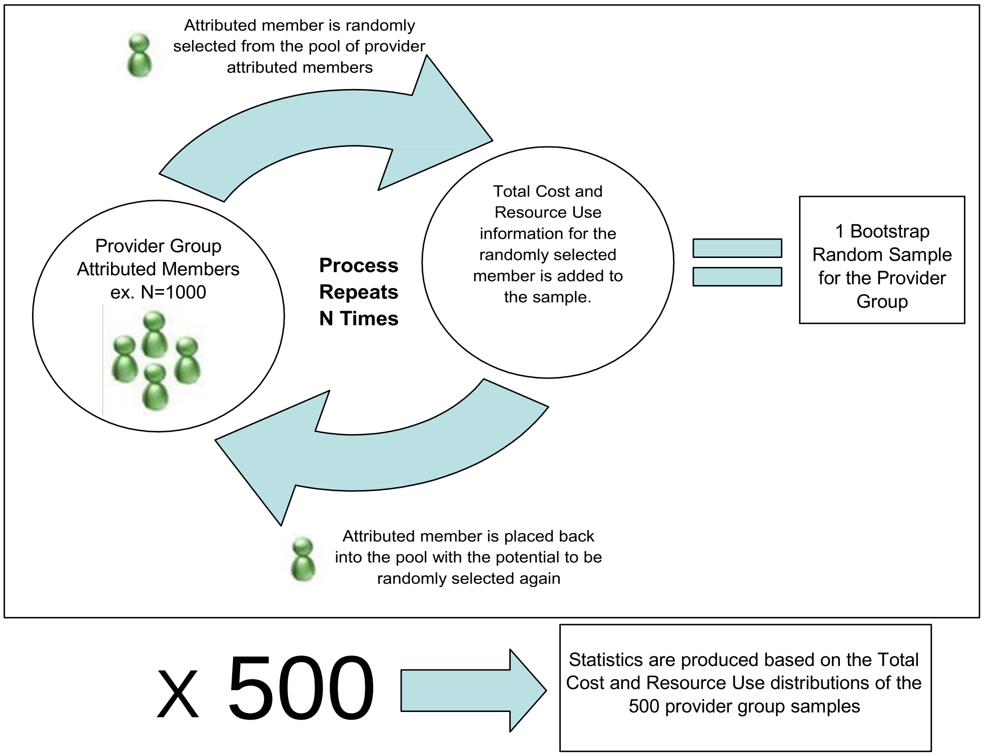
- These differences in provider performance over time occur because of known changes in fee schedules, collaborating provider usage and resource use saving initiatives can account for the differences.
- Since the measure is designed to capture and reflect changes in these areas, we expect to see some explainable variability within a provider group over time.

## *Definitions and Examples*

### Figure 1: 90% Sampling – Simple Random Sample Without Replacement


Provider Group
Attributed Members
ex. N=1000
Attributed member is randomly selected from the pool of provider attributed members
Process Repeats
Until 90% of N is reached
Total Cost and Resource Use information for the randomly selected member is added to the sample
1 Simple Random Sample for the Provider Group
x 500
Statistics are produced based on the Total Cost and Resource Use distributions of the 500 provider group samples

### Figure 2: Bootstrap Sampling – Unrestricted Random Sampling With Full Replacement


Attributed member is randomly selected from the pool of provider attributed members
Provider Group
Attributed Members
ex. N=1000
Process
Repeats
N Times
Total Cost and Resource Use information for the randomly selected member is added to the sample.
1 Bootstrap Random Sample for the Provider Group
Attributed member is placed back into the pool with the potential to be randomly selected again
x 500
Statistics are produced based on the Total Cost and Resource Use distributions of the 500 provider group samples

HealthPartners®

HealthPartners Technical Documentation

# Total Cost of Care and Total Resource Use Validity Testing Analysis

## Purpose

To evaluate the Total Cost of Care and Resource Use measures by comparing the findings and correlations to other known information sources and metrics to determine the validity of the measures.

## Table of Contents

## *Overview of Analysis*

The Total Cost of Care and Resource Use are measures of a provider's risk adjusted cost and resource use effectiveness at managing their primary care attributed population across the care continuum. The Total Cost of Care and Resource Use measures were applied to HealthPartners primary care metro providers as per the specifications of the measures. Additional standard utilization metrics were also applied to the underlying data in the actual and risk adjusted forms. The total cost index (TCI) and total resource use index (RUI) findings are compared by provider group to the actual and risk adjusted utilization metrics to determine the correctness of conclusions.

### Methodology

The Total Cost of Care and Resource Use measures should differentiate between providers based on the cost per member and/or consumption of resources per member given all other factors are equal. The ACG adjustment controls for variations in the illness burden of the patients and the peer grouping controls for various patient demographics, provider types and types of product.  The remaining factors reflect what the provider can control.

The Total Cost of Care and Resource Use measures should show various strengths of correlations to known utilization metrics. These correlation strengths will depend upon how fully encompassing the utilization metric is within the component being measured and whether the metrics are risk adjusted. For example the admit count utilization measure should be highly correlated to the inpatient resource use as the only factor not accounted for in the admit count measure is intensity (aka: level of treatment). When risk adjustment is applied the correlation will be reduced as the illness burden variation is removed.

The Total Cost of Care and Resource Use measures are designed to evaluate the entire patient and/or provider. Since a person centered measure does not currently exist, the utilization metrics are being used as a proxy to evaluate the correctness and accuracy of the conclusions drawn by the Total Cost of Care and Resource Use measures. These comparisons and correlations should be considered as directional and are not absolute. The utilization metrics do not measure intensity or cost per unit and are targeted to measure a specific service therefore the correlations to the Total Cost of Care and Resource Use need interpretation as high correlation are not always the ultimate goal or the expected result.

### Analysis Overview

- The Pearson correlation coefficients are calculated at the network level between provider groups. In general, the correlation coefficient is an indicator of the level of connection or influence two measures have on each another.
- The correlation coefficient scores range from negative one to positive, with the closer to either value indicating the more influence or connection and the close to zero indicating no influence.
- When the correlation is positive both values move in the same direction and when the correlation is negative the values move in the opposite direction.
    - Positive correlation example: the more admits that are incurred, the more total spend is accumulated. In this case the correlation coefficient would be close to 1.0.
- Network Overview Non Risk Adjusted Metrics
    - Correlations between the ACG score and the non-ACG adjusted cost PMPM and TCRRV PMPM.
    - Correlations between known utilization metrics and the ACG score and the non-ACG adjusted cost PMPM and TCRRV PMPM.
    - Correlations between known utilization metrics within specific places of service and the non-ACG adjusted cost PMPM and TCRRV PMPM for the corresponding places of service.  .

- Network Overview Risk Adjusted Metrics
  - Correlations between the ACG score and the Total Cost Index and Resource Use Index.
  - Correlations between known utilization metrics and the overall TCI and RUI.
  - Correlations between known utilization metrics within specific places of service and the TCI and RUI for the corresponding places of service Rx only has a TCI as there is no price variation between providers for pharmacy services.

## Member Population

- Members age 1 – 64 included (babies < 1 and members age 65+ are excluded).
- Members are included if they are enrolled for a minimum of 9 months during the 12 month claims window.
- Commercial products only.
- Attributed members only.
- A member is assigned to the provider group that provides the largest percentage of the primary care office visits.
- In the event of a tie, the provider group with the most recent visit is attributed the member.
- Members that do not have a primary care office visit are excluded from attribution and TCOC.
- Metro Primary Care Providers with more than 600 members that meet the above criteria.

## Network Analysis Overview

- HealthPartners primary care metro network consists of 19 individual provider groups that have 230 clinic sites.
- The total membership of the primary care attributed metro network is over 300,000 members in 2009.
- The variations between provider groups within the following metrics:
  - ACG score variation – 0.85 points (min 0.73 and max 1.59).
  - Total Cost of Care variation – 0.21 points (min 0.84 and max 1.04).
  - Resource Use variation – 0.16 points (min 0.91 and max 1.07).
  - Provider group size vary from 600 to 100,000 members.

## Metrics

- Total Cost Index – TCI: a provider's ACG Adjusted total cost per member per month divided by the metro average ACG Adjusted total cost per member per month.
- Total Care Relative Resource Use Value Index – RUI: a provider's ACG Adjusted total resource use per member per month divided by the metro average ACG Adjusted total resource use per member per month.
  - The Total Care Relative Resource Use Values (TCRRVs) place a relative value unit on all health care services and are the basis of the resource use index (see TCRRV documentation on www.healthpartners.com/tcoc).
- Price Index – PI: a natural byproduct of the TCI and RUI. By definition the only variance between the TCI and RUI is that RUI is void of price.

- Each of these measures is repeated for the four major places of service, inpatient, outpatient, professional and pharmacy.
- Utilization metric indices are counts of distinct services compared to the peer group average.
    - o These utilization metrics are risk adjusted through the ACG methodology, which is accomplished by creating expected value by ACG cell.

## *Overview of Conclusions*

- The Total Cost of Care and the Resource Use measures accurately and consistently identified providers that are low or high performers as the measures were able to evaluate a provider's cost and resource effectiveness as supported by known utilization measures.
- There is a high correlation between ACG score and the unadjusted PMPM and TCRRVs which indicates that the Actual PMPM and the Actual TCRRVs are a good measure of the consumption of resources.
- The ACGs, Actual PMPMs and Actual TCRRVs have similar correlation scores to all utilization metrics which indicate the TCRRVs are performing as expected and are a solid measure of resource consumption.
- The Resource Use measure has a high correlation (0.77) to a composite utilization index, which was developed as a proxy to measure total resource consumption (see RUI vs. Risk adjusted Composite Utilization Index section).
- The Total Cost of Care and Resource Use measures differentiate between provider groups accurately and correctly as supported by a wide array of utilization metrics (see Detailed Provider to Provider Analysis and Detailed Provider to Provider – Selected Place of Service sections).

## *Total Cost of Care & Resource Use Report*

This graphic is displayed for a frame of reference. Each provider group has an ACG index, Total Cost Index and a Resource Use Index and each of these are relative to the metro total. The red line divides providers between above and below the average total cost index. There are also utilization metrics described in the Metric Overview section that are calculated for each provider group that are shown later in the analysis.

### Primary Care Provider Network Overview

Commercial, Continuously Enrolled, Excluding Babies and 65+
Dates of Service within each Year
Indexed to the Metro Average

| | Average ACG Score | | | TCI | | | Resource Use Index | | |
|---|---|---|---|---|---|---|---|---|---|
| **Provider Group** | **2007** | **2008** | **2009** | **2007** | **2008** | **2009** | **2007** | **2008** | **2009** |
| Provider O | 0.98 | 0.96 | 0.96 | 0.83 | 0.86 | 0.84 | 0.91 | 0.95 | 0.91 |
| Provider G | 1.03 | 1.16 | 1.09 | 0.92 | 0.95 | 0.89 | 0.96 | 0.98 | 0.93 |
| Provider M | 1.07 | 1.04 | 1.09 | 0.94 | 0.95 | 0.92 | 1.03 | 1.05 | 1.01 |
| Provider D | 1.02 | 1.03 | 1.03 | 0.99 | 0.88 | 0.93 | 1.00 | 0.92 | 0.95 |
| Provider N | 1.04 | 1.05 | 1.04 | 0.92 | 0.98 | 0.94 | 0.99 | 1.07 | 1.03 |
| Provider F | 1.06 | 1.06 | 1.05 | 0.95 | 0.96 | 0.94 | 1.00 | 1.02 | 1.01 |
| Provider S | 0.94 | 0.92 | 0.92 | 0.91 | 0.93 | 0.95 | 0.97 | 0.99 | 1.01 |
| Provider I | 1.01 | 1.02 | 1.02 | 0.99 | 0.98 | 0.97 | 0.99 | 0.98 | 0.97 |
| Provider Q | 0.90 | 0.92 | 0.97 | 0.94 | 0.99 | 0.98 | 0.97 | 1.02 | 1.00 |
| Provider K | 0.77 | 0.79 | 0.79 | 1.03 | 1.01 | 0.98 | 1.06 | 1.03 | 1.00 |
| Provider L | 0.95 | 0.94 | 0.95 | 0.90 | 0.94 | 0.98 | 0.92 | 0.97 | 1.02 |
| Provider B | 0.93 | 0.94 | 1.00 | 0.95 | 0.90 | 0.99 | 0.98 | 0.94 | 1.04 |
| Provider E | 1.03 | 1.00 | 0.99 | 1.00 | 1.00 | 1.01 | 0.98 | 0.98 | 1.00 |
| Provider R | 1.07 | 1.05 | 1.03 | 0.89 | 0.96 | 1.01 | 0.97 | 1.03 | 1.07 |
| Provider H | 1.01 | 0.96 | 1.00 | 1.02 | 1.05 | 1.04 | 1.04 | 1.07 | 1.06 |
| Provider A | 1.01 | 1.03 | 1.02 | 1.03 | 1.02 | 1.04 | 1.01 | 1.01 | 1.02 |
| Provider C | 0.75 | 0.76 | 0.73 | 1.08 | 1.07 | 1.04 | 1.08 | 1.07 | 1.06 |
| Provider P | 0.96 | 0.95 | 0.94 | 1.07 | 1.05 | 1.04 | 1.03 | 1.01 | 0.99 |
| Provider J | 1.64 | 1.61 | 1.59 | 1.03 | 1.09 | 1.04 | 1.00 | 1.05 | 1.01 |
| **Metro Total** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |

## *Correlations Overview*

### Correlations Between ACG Score, Actual PMPMs, Actual TCRRVs, Risk Adj TCI, and Risk Adj RUI

Since the ACG is an industry standard tool that measures resource use there should be strong correlations between it and the Actual PMPMs and Actual TCRRVs. The Actual PMPM and TCRRV correlations should be similar to the ACG score; however the TCRRV's correlation should be stronger as the Actual PMPMs are not a true unbiased measure of resources, as it is impacted by the unit cost of each of the providers within the analysis.

| **Non-Risk Adjusted** | **Correlation Coefficient** | |
|---|---|---|
| **Metric** | **ACG** | **Actual PMPMs** |
| Actual PMPM | 0.95 | |
| Actual TCRRVs | 0.97 | 0.98 |
| ACG Risk Adj TCI | 0.06 | 0.37 |
| ACG Adjusted RUI | -0.09 | 0.15 |

- There is a high correlation between ACG score and the unadjusted PMPM and TCRRVs which indicates that the Actual PMPM and the Actual TCRRVs are a good measure of the consumption of resources.
- There is a low correlation between ACG score and the risk adjusted TCI and RUI. This would indicate that a provider can have a high or low ACG score and still have a high or low risk adjusted TCI.
- There is a lower correlation between the risk adjusted RUI and Actual PMPMs than the risk adjusted TCI and Actual PMPMs as the risk adjusted RUIs are not impacted by the cost per unit.

## *Non-Risk Adjusted Correlations*

### Correlations Between ACG Score, Actual PMPMs, and Actual TCRRVs to Non-Risk Adj Utilization Metrics

Since the ACG score, Actual PMPMs and Actual TCRRVs are a measure of the consumption of health care services, there should be some correlation between these values and known utilization metrics. These correlations will not be absolute as the utilization metrics encompass only a portion of the total member's experience. It is expected however that the Actual TCRRVs, which is the underlying value that measures resource use, should have similar correlations to the Actual PMPMs and ACG scores.

- The ACGs, Actual PMPMs and Actual TCRRVs have similar correlation scores which indicate the TCRRVs are performing as expected and are a solid measure of resource consumption.
- There is a high correlation between ACG score, Actual PMPM and Actual TCRRVs to the prescriptions per 1,000 and E&Ms per 1,000. Since E&M visits and Rx scripts are a good indicator of member utilization and total health care consumption it is a positive sign that there is a strong correlation to the ACGs, actual PMPMs and TCRRVs.
- The admits per 1,000 and ER per 1,000 have the lowest correlations to the ACG and actual PMPMs which would indicate that these are low volume service and are outcome based measures.

## Correlation Between the Non-Risk Adjusted Place of Service Metrics and Actual PMPMs &Actual TCRRVs

There should be a correlation between the place of service utilization metrics and the Actual PMPMs and TCRRVs of the corresponding place of service. The magnitude of the correlation is dependent upon the utilization metric's penetration within the place of service and the cost and/or resource intensity of the metric. The Actual PMPMs correlation to the utilization metric will also be impacted by the unit cost of each of the providers within the analysis.

## Inpatient Utilization Correlation to the Inpatient Actual PMPMs & Actual TCRRVs:Non-Risk Adjusted

There should be strong correlations between the admit rate to the Actual PMPMs and Actual TCRRVs as the only two factors not measured by the admits are the intensity and unit cost of the services performed.

| Inpatient | Correlation Coefficient | |
|---|---|---|
| Metric | IP Actual PMPMs | IP Actual TCRRVs |
| Admits/1000 | 0.87 | 0.88 |

## Outpatient Utilization Correlation to the Outpatient Actual PMPMs & Actual TCRRVs:Non-Risk Adjusted

There should be solid correlations between the ER and outpatient surgery rates to the Actual PMPMs and Actual TCRRVs as these two utilization metrics combine to encompass approximately 50% of the total outpatient spend.

| Outpatient | Correlation Coefficient | |
|---|---|---|
| Metric | OP Actual PMPMs | OP Actual TCRRVs |
| ER/1000 | 0.85 | 0.78 |
| OP Surgery/1000 | 0.68 | 0.77 |

## Professional Utilization Correlation to the Professional Actual PMPMs & Actual TCRRVs:Non-Risk Adjusted

There should be solid correlations between the E&M visits and Lab/Path services to the Actual PMPMs and Actual TCRRVs as they represent 45% of the professional spend, but they are also are good indicators of patients that consume medical services.

| Professional | Correlation Coefficient | |
|---|---|---|
| Metric | Prof Actual PMPMs | Prof Actual TCRRVs |
| E&M/1000 | 0.77 | 0.80 |
| Lab/Path/1000 | 0.83 | 0.80 |

### Rx Utilization Correlation to the Rx Actual PMPMs: Non-Risk Adjusted

There should be strong correlations between the Rx rates to the Actual PMPMs and Actual TCRRVs as the only factor that is not accounted for in the Rx count metric is the intensity of the drug prescribed. The intensity includes generic usage as well as the variation in cost between drugs.

| Rx | Correlation Coefficient | |
|---|---|---|
| Metric | Rx Actual PMPMs | Rx Actual TCRRVs |
| Rx Count | 0.95 | 0.96 |

## *Risk Adjusted Correlations*

### Correlation Between the Risk Adjusted Place of Service Metrics and TCI and RUI

There should be some correlation between the high cost and resource intensive places of service and utilization measures to the TCI and RUI measures. The low intensive places of service and utilization should have a lower correlation to the overall TCI and RUI measures.

- The TCI is influenced by each provider group's overall cost per unit therefore there should be less correlation to the utilization metrics than the RUI. The following analysis will concentrate on the RUI.
- There is a high correlation between IP RUI and admit rate to the overall RUI.
- The professional RUI has a strong correlation with the overall RUI, while the E&M visits and lab/path services have a low correlation. This would indicate that the remaining professional services have a strong correlation to overall RUI.
- As expected there is no correlation between the Rx TCI and overall RUI as the ACG risk adjustment accounts for the variations in pharmacy usage.
- Both the standard and high tech radiology have some correlation to the RUI.

### Correlation Between Risk Adjusted Place of Service Utilization Metrics and Corresponding TCI and RUI

There should be a correlation between the place of service utilization metrics and the risk adjusted PMPMs and TCRRVs of the corresponding place of service. The magnitude of the correlation is dependent upon the utilization metric's penetration within the place of service and the cost and/or resource intensity of the metric. Since the risk adjustment accounts for variations in illness burden these correlations will be different from their non risk adjusted results displayed in the Correlations Overview section.

## Inpatient Utilization Metric Correlation to the Inpatient RUI – RiskAdjusted

There should be strong correlations between the risk adjusted admit rate and the inpatient TCI and RUI. The only two factors not measured by the risk adjusted admit rate are the intensity and price of the services performed.

- There is a high correlation between the risk adjusted admit rate and the inpatient TCI and RUI. This would indicate that the higher the risk adjusted admit rate the more likely a provider will have a higher than average TCI and RUI.

## Outpatient Utilization Metrics Correlations to the Outpatient TCI and RUI – Risk Adjusted

| Outpatient | Correlation Coefficient | |
|---|---|---|
| Metric | OP TCI | OP RUI |
| ER Cnt | 0.89 | 0.84 |
| OP Surgery | 0.29 | 0.39 |

- There is a high correlation between the risk adjusted ER count and the outpatient TCI and RUI. This would indicate that the higher the risk adjusted ER counts the more likely a provider will have a higher than average outpatient TCI and RUI.
- Outpatient surgery having less of a correlation to the outpatient RUI is an indication that these services are not the driving force behind the outpatient RUI performance.

## Professional Utilization Metrics Correlations to the Professional TCI and RUI – Risk Adjusted

| Professional | Correlation Coefficient | |
|---|---|---|
| Metric | Prof TCI | Prof RUI |
| E&M Visits | 0.41 | 0.46 |
| Lab/Path | 0.57 | 0.37 |

- The professional utilization metrics are moderately correlated to the professional TCI and RUI.
- This result is not unexpected as the professional place of service includes a significant amount of services beyond these two utilization measures (other professional services = 54%).
- It is also not unexpected as having higher than average utilization on diagnostic or management based services does not necessarily indicate a higher resource consuming patient.

## Rx Utilization Metric Correlation to the Rx TCI – RiskAdjusted

| Rx | Correlatio n |
|---|---|
| Metric | RX TCI |
| Rx Count | 0.73 |

- This indicates that more prescriptions equate to a higher Rx TCI, however there is no correlation between the Rx TCI and the overall RUI.

## *Detailed Provider to Provider Analysis*

The Total Cost of Care and Resource Use measure are designed to identify variations between providers accurately and correctly. This section of the analysis will compare findings and results from known utilization metrics to the findings and results from the Total Cost of Care and Resource Use measures. If there are differences in conclusions drawn, the analysis identifies the causes and determines which measure, utilization or Total Cost of Care and Resource Use is more accurate/correct.

Since each utilization metric is designed to measure a portion of health care services, a composite utilization measure is necessary to aide in the evaluation of the accuracy and correctness of the Resource Use measure. Since the TCI includes a cost per unit (price) component, the evaluation is more comparable between the RUI and utilization.

Composite Utilization: A utilization metric was created by weighting each of the underlying utilization metrics by the place of service percent of resources it represents of the total resources.

Composite Utilization Metric =

| | |
|---|---|
| Inpatient | (Admit Rate x 16%) + |
| Outpatient | (average(ER rate, OP Surg Rate, High Tech Rad Rate) x 20%) + |
| Professional | (average (E&M rate, Lab/Path Rate, Std Rad) x 45%) + |
| Pharmacy | (Rx rate x 19%) |

## Primary Care Provider Network Overview

### RUI vs. Risk Adjusted Composite Utilization Index

2009 Commercial, Continuously Enrolled, Excluding Babies and 65+
Indexed to the Metro Average

It is expected that the resources should correlated to the composite utilization metric.

| Provider Group | RUI | Admit | ER Count | OP Surgery | Hightech Rad | E&M | Lab/Path | Std. Rad | Rx Cnt | Composite Utilization |
|---|---|---|---|---|---|---|---|---|---|---|
| Provider O | 0.91 | 0.93 | 0.86 | 0.80 | 0.88 | 0.95 | 0.91 | 0.87 | 1.02 | 0.92 |
| Provider G | 0.93 | 0.51 | 0.82 | 0.96 | 1.02 | 1.05 | 1.08 | 0.90 | 1.07 | 0.93 |
| Provider D | 0.95 | 0.77 | 0.75 | 1.08 | 0.88 | 1.03 | 0.89 | 0.95 | 0.86 | 0.90 |
| Provider I | 0.97 | 0.99 | 0.91 | 0.94 | 0.93 | 0.98 | 1.07 | 1.00 | 0.94 | 0.98 |
| Provider P | 0.99 | 0.97 | 0.95 | 1.14 | 1.06 | 1.03 | 1.10 | 0.94 | 0.95 | 1.01 |
| Provider Q | 1.00 | 1.00 | 1.27 | 1.12 | 0.98 | 1.01 | 0.92 | 0.85 | 1.03 | 1.00 |
| Provider K | 1.00 | 1.19 | 1.23 | 1.17 | 1.07 | 1.00 | 0.86 | 0.85 | 0.93 | 1.00 |
| Provider E | 1.00 | 1.01 | 1.17 | 1.01 | 0.96 | 0.99 | 0.95 | 1.08 | 1.04 | 1.02 |
| Provider S | 1.01 | 1.03 | 1.04 | 1.13 | 1.14 | 0.96 | 0.78 | 1.04 | 1.01 | 0.99 |
| Provider F | 1.01 | 0.92 | 0.87 | 1.00 | 0.97 | 1.02 | 0.87 | 0.85 | 1.03 | 0.94 |
| Provider J | 1.01 | 0.78 | 1.45 | 0.98 | 0.80 | 0.98 | 0.95 | 0.97 | 1.36 | 1.03 |
| Provider M | 1.01 | 1.05 | 0.82 | 1.05 | 0.95 | 0.98 | 0.92 | 1.03 | 1.04 | 0.99 |
| Provider L | 1.02 | 1.05 | 0.89 | 0.86 | 1.41 | 1.04 | 1.09 | 1.11 | 1.00 | 1.05 |
| Provider A | 1.02 | 1.03 | 1.10 | 0.97 | 1.03 | 1.01 | 0.93 | 1.01 | 1.05 | 1.01 |
| Provider N | 1.03 | 0.97 | 0.90 | 1.03 | 1.03 | 0.99 | 1.00 | 0.93 | 1.08 | 1.00 |
| Provider B | 1.04 | 1.01 | 0.58 | 1.09 | 1.01 | 1.07 | 0.98 | 1.14 | 1.04 | 1.02 |
| Provider C | 1.06 | 1.16 | 1.42 | 0.99 | 1.09 | 1.01 | 0.94 | 1.22 | 0.93 | 1.07 |
| Provider H | 1.06 | 0.98 | 1.10 | 1.09 | 1.15 | 1.02 | 0.95 | 1.14 | 1.14 | 1.06 |
| Provider R | 1.07 | 1.10 | 0.75 | 1.02 | 0.94 | 0.98 | 1.07 | 0.94 | 0.97 | 0.99 |
| **Metro Total** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |

- The composite utilization index correlation to overall RUI is 0.77
- The composite index and the RUI have relatively the same index with the exception of provider groups F and R.
    - Provider F's composite utilization metric is 0.94 while their overall RUI is 1.01. The lower than average composite utilization metric is due to the significantly lower than average admit rate, ER services, lab/path and standard radiology services.
        - The professional services are being undervalued due to intensity as the professional RUI is 5% above average (see S12_Sample Score Report).

- Provider R's slightly lower than average composite utilization metric is due to the lower than average ER visits, high-tech radiology, E&M, standard radiology services, and Rx count.
  - The weight of the admit rate in the composite score is undervalued due to intensity as the 10% higher than average admit rate translates to 24% higher than average inpatient resource use (see S12_Sample Score Report).

## High to Low Provider Contrast Analysis

The TCI and RUI should clearly identify providers that are high or low performing and be supported by the risk adjusted utilization metrics.

## Profile of a High Performing Provider (Low TCI and RUI)

- The four top performing providers achieve lower than average resource use with some common markers:
  - Lower than average admit and ER indices.
  - Standard radiology is at or lower than average for all providers.
  - E&M visits are within 5 points of average.
  - All other utilization markers do not have a clear direction.
- The place of service resource use index is near or below average for inpatient, outpatient, and professional components (see S12_Sample Score Report). The Rx TCI is high for one provider, however that provider has extremely low admits, which offset the Rx usage.

## Profile of a Low Performing Provider (High TCI and RUI)

- The lowest performing four providers have some common utilization markers which supports their higher than average resource use:
  - Higher than average in admits or ER or both.
  - These providers have a minimum of one of the other high resource intensive utilization metrics above average.
  - High tech and standard radiology is above average for all but one of the low performing providers. This one exception provider has 10% higher inpatient admissions.
  - E&M visits are relatively around average (one provider is at 1.07).
  - All other utilization markers do not have a clear direction.
- The place of service resource use index is above average for the professional component and at least one of the other 3 components (see S12_Sample Score Report).

## Profile of Providers that do not Fit the Peer Grouping (Excluded from Metro Primary Care Network)

The Total Cost of Care and the Resource Use measures are designed to evaluate providers that are similar in nature and are within the same peer group. Providers that have a significantly different patient mix or patient profile will stand out as outliers.

| | Average ACG Score | | | TCI | | | Resource Use Index | | |
|---|---|---|---|---|---|---|---|---|---|
| **Provider Group** | **2007** | **2008** | **2009** | **2007** | **2008** | **2009** | **2007** | **2008** | **2009** |
| Provider Y | 1.29 | 1.35 | 1.25 | 1.58 | 1.56 | 1.44 | 1.54 | 1.51 | 1.42 |
| Provider Z | 1.32 | 1.14 | 1.21 | 2.11 | 2.26 | 2.03 | 1.47 | 1.52 | 1.40 |

| Provider Group | RUI | Admit | ER Count | OP Surgery | Hightech Rad | E&M | Lab/Path | Std. Rad | Rx Cnt |
|---|---|---|---|---|---|---|---|---|---|
| Provider Y | 1.42 | 1.29 | 1.20 | 1.08 | 2.88 | 1.19 | 1.54 | 1.65 | 1.10 |
| Provider Z | 1.40 | 1.50 | 1.67 | 1.48 | 1.90 | 1.17 | 1.90 | 1.44 | 0.99 |

- Providers Y and Z have significantly higher ACG scores, which in and of itself is not an indication of an outlier provider.
- They also have significantly higher TCIs driven by 40% higher resource use.
- It is known that these providers treat patients that are high users and in need of a complex level of treatment.
- All of the utilization metrics are above average (except for Provider Z's RX count of 0.99)

## Detailed Provider to Provider Analysis – Selected Place of Service

The inpatient admit and the Rx count rates are highly correlated to their place of service RUIs as they encompass the majority of the services within the place of service and the only factors not measured is the unit cost and intensity of service.

## Expanded Inpatient Resource Use vs. Admit Rate Provider Analysis – Risk Adjusted

There is a strong correlation between the risk adjusted admit rate and the risk adjusted inpatient RUI (0.92, see Inpatient Utilization Metric Correlation to the Inpatient RUI section). The only 2 factors not measured by the risk adjusted admit rate are the intensity and price of the services performed. The RUI will account for the intensity of services performed.

- 9 out of 19 groups have lower than average IP RUI
- 1 out of 9 groups had slightly higher than average IP admissions, due to Provider B having a lower than average intensity.
- 10 out of 19 groups have higher than average IP RUI.
- 1 out of 10 groups had a slightly lower than average IP admissions, due to Provider I having more intensive than average admissions.

## Expanded Rx Total Cost Index vs. Rx Count Provider Analysis – RiskAdjusted

There is a strong correlation between the risk adjusted Rx count and the risk adjusted Rx TCI (0.73, see Rx Utilization Metric Correlation to the Rx TCI section). The only factor not measured by the risk adjusted Rx count is intensity (cost per unit is neutral for all providers). Variations in costs of pharmaceuticals and generic rates would express themselves through intensity and be accounted for in the Rx TCI, but not the Rx count metric.

- 9 out of 19 groups have lower than average Rx TCI.
- 4 out of 9 groups have slightly higher than average Rx fills.
- Providers M and O have higher than average percent generic rate which influences the Rx TCI.
- Providers Q and S have slightly lower than average percent generic rate. The higher than average Rx count and lower than average Rx TCI is due to the prescriptions being less resource intensive/costly than average.
- 10 out of 19 groups have higher than average Rx TCI.
- 2 out of 10 groups have lower than average Rx fills.
- Provider R's percent generic rate is 69% compared to the metro average of 74%, which drove their higher than average Rx TCI.
- Provider P had a slightly lower than average percent generic rate. The lower than average Rx count and average Rx TCI is due to the prescriptions being more resource intensive/costly than average.

## Definitions

### Service Category

- Inpatient: Claims on a 1450 claims form and one of the following criteria
    - Room and Board Revenue codes: 100-189, 200-219, 650, 655, 1000-1005
    - Bill Type code: 21, 28, 66, 86
    - Bill Type code of 11 and a revenue code of 190
- Outpatient all other 1450 claim forms
- Professional all 1500 claim forms
- Rx – All pharmacy data

## Total Cost of Care Validity Metric Overview

### Utilization Metrics

Admits An inpatient admission.

ER Count An outpatient claim that includes at least one revenue code between 450-

459. E&M Count E&M CPT codes from a professional claim.

Lab\Path All Laboratory and Pathology CPT codes.

Standard Radiology All radiology CPT codes that are not considered high technology radiology (MRI, CT, nuclear medicine, PET).

Outpatient Surgery All outpatient visits that include one surgical CPT.

High Technology Rad CPT codes from the professional or outpatient place of service that are considered an MRI, CT, nuclear medicine or PET scan. Only one bill is counted if two are submitted for one patient.

Rx Count Script count.

Percent Generic The percent of prescription that are generic.

## Other Metrics

Actual PMPM The actual spend divided by the member months of the population. These are non risk adjusted numbers.

ACG Score At any given level it is the sum of a (member's assigned ACG cell weight x their member months divided by the total member months) of the given level (aka Average ACG weight at any given level).

TCRRV Total Care Relative Resource Value – is a price neutral value that is relative within and across all places of service and types of treatment. In essence it is a standard fee schedule of all services within the health care continuum.

TCRRV PMPM The actual TCRRVs divided by the member months of the population. These are non risk adjusted numbers.

TCI Total Cost Index – the ACG risk adjusted spend PMPM divided by the analysis population's ACG adjusted spend PMPM.

RUI Resource Use Index – the ACG risk adjusted TCRRV PMPM divided by the analysis population's ACG adjusted TCRRV PMPM.

NATIONAL QUALITY FORUM

# COST AND RESOURCE USE MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

**To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return**

## Brief Measure Information

**NQF #:** 2158
**Measure Title:** Medicare Spending Per Beneficiary (MSPB) - Hospital
**Measure Steward:** Centers for Medicare & Medicaid Services
**Brief Description of Measure:** The Medicare Spending Per Beneficiary (MSPB) - Hospital measure evaluates hospitals' risk-adjusted episode costs relative to the risk-adjusted episode costs of the national median hospital. Specifically, the MSPB-Hospital measure assesses the cost to Medicare for services performed by hospitals and other healthcare providers during an MSPB-Hospital episode, which is comprised of the periods immediately prior to, during, and following a patient's hospital stay. The MSPB-Hospital measure is not condition specific and uses standardized prices when measuring costs. Beneficiary populations eligible for the MSPB-Hospital calculation include Medicare beneficiaries enrolled in Medicare Parts A and B who were discharged from short-term acute Inpatient Prospective Payment System (IPPS) hospitals during the period of performance.
**Developer Rationale:** CMS includes the MSPB-Hospital measure within the Hospital VBP program as a measure of efficiency; the Hospital VBP program, however, also provides financial incentives to hospitals based on their performance on additional quality measures. By measuring the cost of care through the MSPB-Hospital measure in combination with these other quality measures, CMS aims to recognize hospitals that can provide high quality care at a lower cost to Medicare.

The MSPB-Hospital measure is designed to promote higher quality care for beneficiaries by financially incentivizing hospitals to improve care coordination, deliver efficient, effective care, and reduce delivery system fragmentation. Specifically, the MSPB-Hospital measure is calculated as the MSPB-Hospital amount compared to the national episode-weighted median MSPB-Hospital amount. This allows hospitals to improve their score by spending relatively less than the episode-weighted median during a given performance period. For instance, hospitals can decrease (i.e., improve) their MSPB-Hospital Amount through actions such as: 1) improving coordination with post-acute providers to reduce the likelihood of hospital readmissions, 2) identifying unnecessary or low-value post-acute services and reduce or eliminate these services, or 3) shifting post-acute care from more expensive services (e.g., skilled nursing facilities) to less expensive services (e.g., home health) in cases that would not affect patient outcomes.

Care coordination helps ensure a patient's needs and preferences for care are understood, and that those needs and preferences are shared between providers, patients, and families as a patient moves from one healthcare setting to another. People with chronic conditions, such as diabetes and hypertension, often receive care in multiple settings from numerous providers. As a result, care coordination among different providers is required to avoid waste, over-, under-, or misuse of prescribed medications and conflicting plans of care.
**Resource Use Measure Type:** Cost per episode

**Data Source:** Claims (Only); Other
**Level of Analysis:** Facility
**Costing Method:** Standardized prices
**Tested Population:** Medicare

**Attribution Approach:** MSPB-Hospital episode is attributed to the hospital on the trigger inpatient claim for the index hospital admission that begins an MSPB-Hospital episode. Hospitalizations eligible to start an MSPB-Hospital episode must end in a discharge 30 days prior to the end of the period of performance to permit the collection of claim information during the post-discharge period. Exceptions: Acute-to-acute transfers during the index admission are not considered index admissions for the purposes of the MSPB-Hospital measure. Neither the transferring hospital nor the receiving hospital will have an index admission attributed to them.

**Risk Adjustment:** Ordinary least squares (OLS) linear regression model based on the Centers for Medicare & Medicaid Services' hierarchical conditions categories (CMS-HCC) risk adjustment methodology. Independent variables included in the model: beneficiary age, health status (measured by hierarchical condition categories (HCCs), disability status, end-stage renal disease (ESRD) status, resident in a long-term care facility, MS-DRG indicators for the index admission, and disease interactions (HCCs x enrollment status).

**IF Endorsement Maintenance – Original Endorsement Date:** Dec 09, 2013 **Most Recent Endorsement Date:** Dec 09, 2013

## Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement, endorsed measures are evaluated periodically to ensure that the measure still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

### Criteria 1: Importance to Measure and Report

**1a. High Priority**
**Maintenance measures – less emphasis on this criterion unless there is new information or change in evidence since the prior evaluation.**

**1a. High Priority**. This requirement involves demonstrating that the measure focus addresses one of the following:

- A specific national health goal/priority identified by the Department of Health and Human Services or the National Priorities Partnership convened by NQF.
- A demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality).

**Summary of information provided to fulfill the High Priority requirement**

- To demonstrate this measure focuses on a high-priority area, the developers cite data indicating Medicare expenditures accounted for 3.6% ($647.6 billion) of the Gross Domestic Product (GDP) in 2015 and hospital benefits accounted for 30% ($188.3 billion) of those Medicare expenditures. The developer also cites data indicating Medicare expenditures will account for 6.0 to 9.1% of the GDP by 2090, if current trends continue.

**Preliminary rating for High Priority:** ☒ High ☐ Moderate ☐ Low ☐ Insufficient

**1b. Gap in Care/Opportunity for Improvement** and **1b. Disparities**
**Maintenance measures – increased emphasis on gap and variation**

**1b. Performance Gap.** This requirement involves demonstrating a resource use or cost problems exist and there is an opportunity for improvement (i.e., data demonstrating variation in the delivery of care across providers and/or population group (disparities in care)).

- The developer provided performance data from one period of performance - January 1, 2015 to December 31, 2015. This data represents 4.2 million Medicare beneficiaries and 3,298 inpatient prospective payment system hospitals with at least 25 episodes for the performance period.

- This hospital level measure calculates the ratio of payment standardized, risk-adjusted Medicare Spending Per Beneficiary (MSPB) amount for each hospital divided by the episode-weighted median MSPB-Hospital amount across all hospitals. Lower scores are better. 2015 performance data are provided in the table below. Measure scores ranged from 0.59 to 2.25 with an interquartile range of 0.09. These values indicate performance variation among providers.

| | |
|---|---|
| Mean (SD) | 0.99 (0.09) |
| Range | 0.59 – 2.25 |
| 10th percentile | 0.89 |
| 20th percentile | 0.92 |
| 25th percentile | 0.94 |
| 30th percentile | 0.95 |
| 40th percentile | 0.97 |
| Median | 0.99 |
| 60th percentile | 1.00 |
| 70th percentile | 1.02 |
| 75th percentile | 1.03 |
| 80th percentile | 1.04 |
| 90th percentile | 1.08 |
| Interquartile Range | 0.09 |

- To examine changes in performance over time, the developers calculated score changes between 2014 and 2015. Between these two years, measure scores improved (i.e., decreased) for 47.46% of hospitals. A summary of the changes over time is provided in the table below.

| Performance by decile | Percentage Change from 2014 to 2015 (Lower is better) |
|---|---|
| 10th percentile | -4.08% |
| 20th percentile | -2.30% |
| 30th percentile | -1.25% |
| 40th percentile | -0.47% |
| Median | 0.16% |
| 60th percentile | 0.84% |
| 70th percentile | 1.66% |
| 80th percentile | 2.83% |
| 90th percentile | 4.99% |

**1.b Disparities.**

- To examine disparities by population group, the developers examined the effect of socioeconomic status (SES) and sociodemographic status (SDS) on measure scores in the risk adjustment model. SES was captured via an income-to-poverty ratio for each 5-digit zip code. This ratio was created using 5-year estimates data from the American Community Survey. For each beneficiary, an income-to-poverty ratio was estimated using the beneficiary's 5-digit zip code. Beneficiary race (i.e., Black or Non-Black) was determined using data from the Medicare Enrollment Database.
- When the SES (i.e., income-to-poverty ratio) and SDS (i.e., race) variables were included in the measure's risk adjustment model, changes in hospitals' measure scores were minimal for the majority of hospitals.

| Variable | Effect on Measure Score |
|---|---|
| Income-to-Poverty Ratio | ± 0.01 or less for 97% of hospitals |
| Race | ± 0.01 or less for 95% of hospitals |

***Questions for the Committee:***

- *Is there a gap in care that warrants a national performance measure?*
- *Are the variables included in the disparities analysis appropriate? Are the conclusions made by the developer on excluding the SDS factors appropriate?*

**Preliminary rating for opportunity for improvement: ☒ High ☐ Moderate ☐ Low ☐ Insufficient**

## 1c. Measure Intent

**1c. Intent of the cost or resource use measure.** This requirement involves describing the measure intent of the resource use measure and the measure construct.

- The developer states the measure's intent is to, "…incentivize hospitals to coordinate care and reduce unnecessary utilization during the period immediately prior to, during, and in the 30 days after a hospital discharge."
- The developer describes the measure construct as encompassing all types of services received (i.e., Part A and Part B claims) during the episode and states that the all-cause nature of the measure maximizes its ability to promote hospital efficiency by promoting coordination across settings and providers.

***Questions for the Committee:***

- *Is the measure clearly described?*
- *Is it appropriate to measure costs or resource use in this way for this condition? In this care setting? At this level of analysis?*
- *Are the costs included appropriate and consistent with the measure intent?*
- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

**Preliminary rating for measure intent: ☒ High ☐ Moderate ☐ Low ☐ Insufficient**

**Committee pre-evaluation comments**
**Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)**

***1a. High Priority***

Comments:
**This is a high priority measure
**Agree with high priority.
**Yes, the measure is high priority
**Resource use across acute care episodes is an important topic, especially because of the wide variation in utilization of post-acute care services.
**Global (non-disease specific) cost measure for hospitalized Medicare patients. Hospital care accounts for a substantial portion of health care costs.
**Yes, measure meets sub-criterion. NQF assessment captured adequately.
**This is a high priority area. A concern I have is looking at resource and cost measures absent the disease states where the measure will be operationalized. There is a link, and the NQF building blocks speak to this approach. The intersection of cost and quality, not cost as a stand-alone construct.
Efficiency is defined by NQF as the resource use (or cost) associated with a specific level of performance. These stand alone cost measures create incentives for providers and organizations to cut utilization first, in an attempt to bring down cost scores.
**Yes, high priority--cost of care/affordability and reducing variation. A key focus of HHS to improve care coordination to improve quality and reduce costs.
**The AAMC strongly disagrees with the NQF's "Preliminary rating for opportunity for improvement," which is rated as high.
In the worksheet, NQF provides the following methods for improving on this measure: "1) improving coordination with post-acute providers to reduce the likelihood of hospital readmissions, 2) identifying unnecessary or low-value post-acute services and reduce or eliminate these services, or 3) shifting post-acute care from more expensive services (e.g., skilled nursing facilities) to less expensive services (e.g., home health) in cases that would not affect patient outcomes."

While AMCs are addressing these critical issues, these recommendations require substantial effort and buy-in from multiple hospital units and the post-acute care community. The measure covers spending across a multitude of conditions and DRGs, making targeted interventions extremely difficult. Given the complexity of the measure, there are additional concerns that the measure may be capturing variation that is beyond the hospital's direct control. Even significant efforts to address the recommendations outlined may not be enough to "move the needle."

***1b. Performance Gap***

Comments:

**Total spending variation represents a potential care disparity.

**Data indicate wide variation in performance.

**There is variability, with the interquartile range varying from 0.94-1.03 and the 10-90% range varying from 0.89-1.08, or $4000 on an average of $20,000. The risk adjustment controls for DRG and prior status in a SNF, so the principal sources of variation in this measure are probably (we are not presented with this information, althought hospitals are provided info in their reports) readmissions and home health. Disparities are only discussed in the context of SDS and SES risk adjustment, and we don't have data on how hospitals with high levels of low SES/SDS patients fare compared to those with low levels.

**There is variation across hospitals on MSPB with the 10th and 90th percentiles being about 10% below and above the median hospital.

**Yes, I agree that a gap in care exists. Regarding disparities, the developers considered SES and SDS, but not regional and rural vs urban disparities in care.

**Yes, measure meets sub-criterion.

**There is a gap, and the measure appears to positively address through data and opportunity to close the gap - or at a minimum reduce the delta between years.

**Yes, there is variation across hospitals in MSBP (mostly due to variation in SNF component). Unclear why they didn't test using a beneficiary's dual status as measure of low income (i.e., dual at any point in the year). For 3% of the hospitals it seems to matter; unclear how much it shifted the percentile rank of these facilities.

## Criteria 2: Scientific Acceptability of Measure Properties

### 2a. Reliability

### 2a1. Reliability Specifications

**Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures**

**2a1. Specifications.** This requirement involves providing the full specifications for the measure so that it can be implemented consistently within and across organizations and allow for comparability. Electronic health record (EHR) measure specifications are based on the quality data model (QDM).

**Data source(s):**

- Medicare Part A & B claims
- Medicare Enrollment Database (EDB)
- Minimum Data Set (MDS) (risk adjustment model)
- American Community Survey (ACS) (used to evaluate the inclusion of SES/SDS in the risk adjustment model)

**Specifications:**

- This hospital level measure calculates the ratio of payment standardized, risk-adjusted Medicare Spending Per Beneficiary (MSPB) amount for each hospital divided by the episode-weighted median MSPB-Hospital amount across all hospitals. Lower scores are better.
- The numerator includes the average spending level for the hospital's MSPB-hospital episodes divided by the average expected episode spending level for the hospital's episodes, multiplied by the average spending over all episodes across all hospitals nationally.
- The denominator includes the episode-weighted median MSPB-Hospital amount across all episodes nationally.
- The measure's construction logic contains eight steps:
  - Step 1 Standardize claim payments: Using the Centers for Medicare & Medicaid Services' (CMS)

payment standardization methodology, the developer standardizes claims payments to account for payment variation related to local or regional price differences and add-on or incentive adjustments.

  - Step 2 Calculate standardized episode spending: Episode spending is the sum of the standardized Medicare claims payments for the episode (i.e., 3 days prior to hospitalization + length of hospital stay + 30 days post-hospital discharge).
  - Step 3 Calculate expected episode spending: The effect of select beneficiary factors (e.g., age, health status, enrollment status) on cost is first estimated via ordinary least square (OLS) regression. Using the values from the OLS regression, expected spending is then calculated for each major diagnostic category (MDC) in a multivariate regression.
  - Step 4 Winsorize predicted values: To mitigate the effect of extreme predicted values, predicted values in the 0.5th percentile are "bottom-coded" and the predictive values distribution is renormalized.
  - Step 5 Calculate residuals: To estimate the relationship between standardized episode spending and expected episode spending, residuals are calculated as the difference between the two.
  - Step 6 Exclude outliers: At the episode level, residuals falling above the 99th percentile or below the 1st percentile of the residual distribution across all episodes are excluded. This mitigates the effect of high-cost and low-cost outliers.
  - Step 7 Calculate MSPB-Hospital amount for each hospital:

$$\left(\frac{Average\ Standardized\ Episode\ Spending}{Average\ Expected\ Episode\ Spending}\right) \times \left(\begin{matrix} Average\ Standardized\ Spending \\ Across\ All\ Episodes \end{matrix}\right)$$

  - Step 8: Calculate the MSPB-Hospital measure:

$$\left(\frac{MSPB\ Hospital\ Amount}{Episode\ Weighted\ National\ Median\ MSPB\ Amount}\right)$$

- The measure's clinical logic describes the measure's grouping methodology, cost calculation, measure trigger and end mechanisms and co-morbid and interactions. The purpose of these steps is to create a coherent cohort of beneficiaries for whom accurate episode costs can been estimated.
- Adjustments for Comparability: The measure developer used the following inclusion, exclusion, and risk adjustment approach to account for patient severity.
  - Included populations:
    - Medicare beneficiaries enrolled in Medicare Parts A and B admitted to subsection (d) hospitals, defined as "hospitals in the 50 States and D.C. other than: psychiatric hospitals, rehabilitation hospitals, hospitals whose inpatients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer".
  - Excluded populations:
    - Episodes for which the index admission inpatient claim has $0 actual payment or a $0 standardized payment
    - Episodes during which the beneficiary was transferred to another acute care hospital
    - Admissions to hospitals that Medicare does not reimburse through the Inpatient Prospective Payment System (IPPS) system
    - Episodes whose relative scores fall above the 99th percentile or below the 1st percentile of the distribution of residuals
    - Episodes for which full data are not available:
      - Beneficiary only enrolled in Medicare Part A
      - Beneficiary becomes deceased
      - Beneficiary enrolled in Medicare Advantage or had Medicare as a secondary payer any time 90 days before or during the episode
      - Beneficiary's primary insurance becomes Medicaid during an episode

- Risk adjustment
  - The measure is risk adjusted for age and severity of illness. The independent variables within the model:
    - Beneficiary age
    - Health status – measured via hierarchical conditions categories (HCCs)
    - Disability status
    - End-stage renal disease (ESRD) status
    - Residence in a long-term care facility
    - MS-DRG indicators
  - All the independent variables are calculated using Medicare claims data during the 90 days prior to the start of an episode
  - The risk adjustment approach uses an ordinary least squares (OLS) linear regression model that is stratified by the index admission's MDC. This approach is similar to the CMS-HCC risk adjustment methodology, though it should be noted  the MSPB-Hospital risk adjustment approach does not include sex as an independent variable.

***Questions for the Committee***:

- *Are all the data elements clearly defined?  Are all appropriate codes included?*
- *Is the clinical logic clear? Is the construction logic clear?*

**2a2. Reliability Testing  Testing attachment**
**Maintenance measures – less emphasis if no new testing data provided**

**2a2. Reliability testing:** This requirement involves demonstrating that the measure data elements are repeatable, produce the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

**For maintenance measures, summarize the reliability testing from the prior review:**

- In the 2012 submission, the developer provided a summary of the data element and measure score reliability testing conducted using data obtained between 5/15/2010 and 2/14/2011. To demonstrate data element reliability, the developer cited CMS's extensive auditing program. To demonstrate measure score reliability, the developer used four approaches: (1) Test/Retest, (2) Seasonality, (3) Reliability Score, and (4) Bootstrapping.
- In the 2012 review, the Committee passed the measure on reliability (voting results: High-10; Moderate-14; Low-1; Insufficient-0), but raised a concern that the test/retest results showed approximately 30% of hospitals in the lowest spending quintile in one sample were not in the lowest spending quintile in the next sample and 30% of hospitals in the highest spending quintile in one sample were not in the highest spending quintile in the next sample. The developer cited high Spearman rank correlation coefficient for a hospital across samples ($\rho$=0.835) as an indicator that using a different random group of patients does not result in significant variation of the hospital's relative performance.

**Describe any updates to testing:**

- For this maintenance submission, the developer tested data element and measure score reliability using data from approximately 5.5 million episodes that occurred between 1/1/2015 and 12/1/2015.
  (See testing details below).

**SUMMARY OF TESTING**

**Reliability testing level** ☐ **Measure score** ☐ **Data element** ☒ **Both**

**Reliability testing performed with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**

**Method(s) of reliability testing**

Updated Testing

- To test reliability, the developer used data obtained from 3,298 IPPS hospitals between 1/1/2015 and 12/1/2015.
- *Data element reliability*

- To demonstrate data element reliability, the developer cited CMS auditing and data analysis programs that regularly assess the accuracy of the claims submitted to CMS. To enhance the reliability of the data elements, the measure is calculated using data with a 3 month claims run-out from the end of the performance period.

- *Measure score reliability*
    - To demonstrate measure score reliability, the developer conducted two analyses:
        1. Test/Retest analysis: a similar approach was used as in the initial testing, but the developer compared two random subsets from 2015, and compared the set of 2015 episodes to the set of 2014 episodes.
        2. Reliability score: the developer used a similar approach to calculate reliability scores.

**Results of reliability testing**

Updated Testing

- *Data element reliability*
    - The developer did not provide results to demonstrate data element reliability.
- *Measure score reliability*
    - Test/Retest analysis:
        - 2015 vs. 2014 measure scores: over 75% of hospitals in the lowest-spending quintile in one year were in the same quintile in the other year; over 74% of hospitals in the high-spending quintile in one year were in the same quintile in the other year. Spearman rank correlation coefficient for a hospital across the two years was 0.85 and the Pearson correlation coefficient was 0.81, both indicating a high degree of agreement between the two years.
        - 2015 random subset$_1$ vs. 2015 random subset$_2$: over 72% of hospitals in the lowest-spending quintile in one subset were in the same quintile in the other subset; over 71% of hospitals in the highest-spending quintile in one subset were in the same quintile in the other subset. Spearman rank correlations for a hospital across samples was 0.82, and the Pearson correlation coefficient was 0.70. The developer states this lower value for the Pearson correlation coefficient is acceptable given the outcome of interest (i.e., measure scores) is identical in the two subsets and this negatively affects the calculation of the correlation coefficient.
    - Reliability score calculations:
        - For hospitals with at least 25 MSPB-Hospital episodes, over 99% had a reliability score greater than 0.4 and 67.9% had a reliability score greater than 0.9. The developer cites previous work supporting 0.4 as the lower limit of moderate reliability.

***Questions for the Committee:***

- *Does the Committee agree that citing CMS auditing and data analysis programs is an adequate demonstration of data element validity?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the measure score reliability results demonstrate sufficient reliability so that differences in performance can be identified?*

**Guidance from the Reliability Algorithm**

**Precise specifications (Box 1) → Empiric reliability testing (Box 2) → Score-level testing (Box 4) → Appropriate method (Box 5) → Moderate certainty that measure results are reliable (Box 6b)**

**Preliminary rating for reliability: ☐ High ☒ Moderate ☐ Low ☐ Insufficient**

**2b. Validity**

**Maintenance measures – less emphasis if no new testing data provided**

**2b1. Validity: Specifications**

**2b1. Validity Specifications:** This requirement involves demonstrating that the measure specifications are consistent with the measure intent described under criterion 1c and capture the most inclusive target population.

**Specifications consistent with intent described in 1c.** ☒ **Yes** ☐ **Somewhat** ☐ **No**

***Question for the Committee:***

- *Does the Committee agree the specifications are consistent with the intent of the measure?*
- *Is the attribution approach consistent with the measure intent?*
- *Does the accountable entity have reasonable control over the resources measured?*

## 2b2. Validity testing

**2b2. Validity Testing** This requirement involves demonstrating that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided.

**For maintenance measures, summarize the validity testing from the prior review:**

- In the 2012 submission, the developers tested validity by correlating the measure score with the percent of beneficiaries with multiple episodes and other outcomes measures specifically, heart attack, heart failure, and pneumonia readmission rates. The same data used in the initial reliability testing was used in the initial validity testing.
- In the 2012 review, the Committee passed the Validity criterion on the measure (voting results High- 0; Moderate-13; Low-11; Insufficient-1), but raised concerns about the construct validity testing results, which demonstrated low correlation with measures of readmissions in heart attack, heart failure, and pneumonia, the length of the look back period for the HCC risk adjustment model, and the appropriateness of not incorporating the dual eligible population into the risk adjustment model.

**Describe any updates to validity testing:**

- For this maintenance submission, the developer conducted additional validity testing by examining the measure score's correlation with other measures of spending and service utilization and examining cost variation by time period. The same data used in the updated reliability testing was used in the updated validity testing.

**SUMMARY OF TESTING**

**Validity testing level** ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

**Method of validity testing of the measure score:**

☐ **Face validity only**

☒ **Empirical validity testing of the measure score**

**Validity testing method:**

Updated Testing

- The developer conducted three new analyses to test the validity of the measure score. These analyses were:
  1. correlation between the MSPB-Hospital measure and the measure of risk-adjusted, aggregated annual per-capital spending for all Medicare beneficiaries produced by CMS at the Hospital Referral Regions (HRR) level. The developer calculated these correlations for the years 2007-2014;
  2. correlation between the MSPB-Hospital measure and a measure of service utilization, calculated as hospital-level averages of services billed during the MSPB-Hospital episode across various categories (e.g., evaluation and management, post-acute, etc.); and
  3. examination of cost variation by time period (i.e., 3 days prior to index admission, index admission length of

stay, and the period post-discharge).

**Validity testing results:**

Updated Validity Testing

- The results from the new validity analyses were:
  1. Correlation with the corresponding HRR-level measure:
     - For each year, the MSPB-Hospital measure had a moderate or strong positive correlation of at least 0.5 with the corresponding HRR-level of measure. The range for the Spearman rank correlation coefficient was 0.53 to 0.63 and the range for the Pearson correlation coefficient was 0.51 to 0.61.
  2. Correlation with Service Utilization
     - The Pearson correlation coefficient between the MSPB-Hospital measure and professional E&M services was 0.42 and 0.52 for the correlation between the MSPB-Hospital measure and post-acute skilled nursing and inpatient service per episode. These results demonstrate a moderate positive correlation between the MSPB-Hospital measure and professional E&M services, post-acute skilled nursing, and inpatient services.
  3. Cost Variation by Time Period
     - In line with the developer's expectations, the post-discharge period accounted for over 84% of total variance in the measure score. The 3 days prior to the index admission and the index admission length of stay accounted for 11% of total variance in the measure score.

***Questions for the Committee:***

- *Do the results demonstrate sufficient validity so that conclusions about resource use can be made?*

## 2b3-2b7. Threats to Validity

**2b3. Exclusions**: This requirement involves demonstrating that the exclusions are:

- supported by the measure intent

AND/OR

- There is a rationale or analysis demonstrating that the measure results are sufficiently distorted due to the magnitude and/or frequency of the non-clinical exclusions;

AND

- Measure specifications for scoring include computing exclusions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion);

AND

- Patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

**Summarize approach and analysis of exclusions**

- Episodes in the following categories are excluded from the measure:
  - Acute-to-acute transfer episodes: based on claim discharge code
  - Death episodes: beneficiary dies during the measurement episode
  - Overlapping episodes: occurrence of an inpatient admission during the 30 days post-discharge of an index admission is not considered a new index admission
  - Outlier episodes: episode whose relative scores fall above the 99th percentile or below the 1st percentile of the distribution of residuals
- To examine the impact of these exclusions, measures scores were recalculated with excluded episode type included in the calculation.
- The results (see summary below) of these analyses indicated the exclusions had minimal impact on the measure score as demonstrated by the effects of the measures scores and the high correlation coefficients

between the measure score calculated with the episode type excluded and the measure score with the episode type included.

| Episode Type | % of total episodes | Effect on Measure Score | Correlation between measure scores |
|---|---|---|---|
| Acute-to-Acute Transfer | 1.6% | 81% change less than ± 0.03<br>>2% change by more than ± 0.10 | 0.95 |
| Death | 8% | 96% change less than ± 0.03<br>>0.2% change by more than ± 0.10 | 0.99 |
| Outlier | N/A | 6% change by more than ± 0.03<br>≈2% change by more than ± 0.10 | 0.93 |
| Overlapping | 12% | 97% change less than ± 0.03<br>0.4% change by more than ± 0.10 | 0.99 |

N/A = Not available

***Questions for the Committee:***

- o *Are the exclusions consistent with the intent of the measure?*
- o *Are the percentages of exclusions what one would expect for each of the exclusion categories?*
- o *Are carve-outs appropriately addressed?*
- o *Are any patients or patient groups inappropriately excluded from the measure? Does the Committee agree with the exclusions for overlapping episodes.*

2b4. Risk adjustment: This requirement involves specifying an evidence-based risk-adjustment strategy (e.g., risk models, risk-stratification) that is based on patient clinical factors that influence the measured outcome and are present at the start of care and has demonstrated adequate discrimination and calibration. If a risk adjustment strategy is not provided, a rationale or data to support no risk-adjustment/-stratification must be provided.

**Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

**Conceptual rationale for SDS factors included ?** ☒ **Yes** ☐ **No**

**SDS factors other than age included in risk model?** ☐ **Yes** ☒ **No**

**Risk adjustment summary**

- The MSPB-Hospital risk adjustment model is based on the CMS-HCC risk adjustment methodology, but unlike the CMS-HCC methodology, the MSPB-Hospital model does NOT adjust for sex.
- The measure employs an ordinary least squares (OLS) regression model and a separate OLS regression model to obtain the predicted episode cost for each Major Diagnostic Category that is determined by the MS-DRG of the index hospital stay.
- The MSPB-risk adjustment model includes indicators of age, disability status, end-stage renal disease status, long-term care, severity of illness (measured via hierarchical conditions categories (HCC)), and the MS-DRG of the index admission.

**Empirical Summary of SDS**

- Race (i.e., Non-Black and Black) and income-to-poverty ratio were used to examine the impact of SDS on the risk adjustment model. Three analyses were conducted:
  - o F-test of significance:
    - ▪ An F-test of significance allows one to see whether the addition of a variable to a regression model has a significant effect on the outcome variable. Both race and income-to-poverty ratio were significant predictors of the measure score, but when included in the risk adjustment regression with other variables, minor change occurred in the measure score.

- Differences in MSPB-Hospital Measure scores

| Variable | Effect on Measure Score |
|---|---|
| Income-to-Poverty Ratio | ± 0.01 or less for 97% of hospitals |
| Race | ± 0.01 or less for 95% of hospitals |

- Correlation between MSPB-Hospital measure scores calculated *with* SES or SDS variable and MSPB-Hospital measure scores *without* the SES or SDS variable

| Variable | Correlation Coefficient |
|---|---|
| Income-to-Poverty Ratio | >0.998 |
| Race | >0.997 |

- The developers stated that the minimal effect of these two variables likely indicates SDS effects on measure scores are largely captured through existing risk adjustment variables and their inclusion in the risk adjustment model is not necessary.

**Risk Model Discrimination and Calibration**

- For discrimination testing, the developer calculated the average R-squared value across all MDCs and the overall R-squared, which is the difference between the observed costs and the national mean cost across all MDCs. A R-squared value represents how close the data are to the fitted regression line. The R-squared values for these were 0.3014 and 0.4757, respectively. These values indicate approximately 30% of the variation in the cost across all MDCs is explained by the risk model and approximately 48% of the variation in observed costs in explained by the risk model.
- For calibration testing, the developer examined options for various lengths of the look-back period and options for stratification of the risk adjustment model.
  - Changing the look back period from 90 to 365 days resulted in the loss of 6.7% of episodes and decline in the overall model fit (i.e., average of R-squared across all MDCs) from 0.3014 to 0.2997.
  - Adding an indicator of institutional status to the stratification plan resulted in minimal improvement in the model's R-squared value and decreased the number of variables that were statistically significant predictors in the model.
- To examine the validity of the model, the developers calculated predictive ratios by risk deciles. Results indicated the model is consistent in predicting spending in all deciles.

***Questions for the Committee:***

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Is the look-back period used for the measures risk-adjustment strategy appropriate?*
- *Are all of the risk adjustment variables present at the start of care?*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors in their risk-adjustment model?*

2b5. Meaningful difference: This requirement involves demonstrating, through data analysis, that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically meaningful differences in performance.

- To demonstrate the measure's ability to identify meaningful differences, the developer stratified measure scores by hospital characteristics and compared the results to expected findings discussed in the literature.

- Results indicated the measure was able to detect differences among hospitals by geographic region, teaching hospital status, and location (i.e., rural versus urban).

***Question for the Committee:***

o *Does this measure identify meaningful differences about cost or resource use?*

2b6. Comparability of data sources/methods: This requirement involves demonstrating that if multiple data sources/methods are specified, they produce comparable results.

N/A

2b7. Missing Data: This requirement involves describing how missing data are handled and demonstrating that the presence of missing data does not bias the measure.

- The developer states that all required data are readily available and retrievable. Missing data does not appear to be an issue for this measure.

**Guidance from the Validity Algorithm**

Precise specifications (Box 1) → Empirical testing conducted with measure as specified (Box 2) → Score-level testing conducted (Box 4) → Method of testing appropriate (Box 5) → moderate certainty that the scores are reliable

**Preliminary rating for validity:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b)**

***2a1. & 2b1. Specifications***

***2a1. Reliability Specifications***

Comments:

** The episode definitions are the key to this measure. Basing performance on DRGs limits the analysis to these coarse measures. Many clinical elements may be inadequately captured. Are all of the episodes specifically DRG? Or will the other episodes of care that are part of the QPP be incorporated into this measure?
** Claims based measure; logic appears straightforward. Reliability is not a major concern.
** Methods haven't changed since initial submission. Measure can be consistently implemented. Methodology is well specified and has been implemented repeatedly.
** Based on claims data - no issues with data elements. Complex process to obtain the measure, but the steps are well presented.
** Data elements are adequately defined.
** Yes, measure meets sub-criterion. NQF assessment captured adequately.
** Comfortable with the measure - and that it can be implemented consistently
**Moderate: concern about the disability variable drawn from enrollment file. Better disability indicator comes from CMS Integrated Data Repository (OREC) variable, which manages to continue to code someone as disabled once they age into Medicare. The enrollment variable converts at age 65 as aged in, so you are missing some disabled folks.

***2a2. Reliability Testing***

Comments:

** It seems that reliability testing between facilities revealed consistent results year to year. Does this represent patient population or facility behavior? I.E. are tertiary care facilities consistently higher because of case complexity or because of inefficiencies?
** Reliability testing consistent with methods used at time of endorsement.
** Reliability testing is sufficient for application to Medicare patients
** Yes. Additional reliability testing presented since last endorsement. Conducted at measure score level. At the data level, the developer using auditing programs to assess accuracy of data but did not provide results from reliability analyses.
** In reviewing the content, I a comfortable with the CMS auditing program - and the demonstration of empiric reliability.
**Moderate to low. Average reliability is .879 which is good, however, the range used by the measure developer runs down to 0.4 which is considered low. Usually the threshold for high stakes applications is 0.70 or higher, and sometimes it will be dropped to 0.65 for QI feedback. I'm concerned about the 0.4 threshold for discriminating differences between hospitals. What fraction of hospitals have reliability in range from 0.4 and 0.7?

***2b1. Validity Specifications***

Comments:

** The specifications seem reasonable.

** Specifications are consistent with intent. Continued questions of extent to which hospitals can influence costs beyond the hospitalization that contribute to variation in this measure. Use of standardized pricing and control for DRG and pre-hospitalization institutional status mask variation in resources actually employed at hospital level in treatment.

** No issues, attribution is good. Exclusion of both transferring and receiving hospitals seems appropriate. How are bundled payments treated?

** In reviewing my material from the previous vote (mid-range), I did not find anything (analyses) that would necessarily improve my vote, nor lower its value. I believe the approaches used in 2B and 2A were consistent with that discussed in 2012.

**High--specifications seem OK save for the disability indicator.

***2b2. Validity Testing***

Comments:

** Developer describes a new approach to testing validity, by comparing the metric to output of Hospital Referral Region values. Correlations there are moderate. I do not see where these approaches to validity testing have been established as standards. I would like to know why the Developer picked this, as opposed to looking at specific conditions as with the earlier submission.

** Committee should walk through validity testing.

** Validity testing consistent with methods used at time of initial endorsement.

** Validity testing indicates that MSPB has moderate correlation with regional population-based spending and with other utilization measures. Most of the spending variation (84%) occurs during the post-discharge timeframe. With MSDRG standard payments for the hospital and incorporation of MSDRG in the expected values, this is not greatly surprising.

** The correlation coefficients reported are only moderate, implying that a significant portion of the variability of the results is not accounted for by the variables measured.

** MSPB captures costs 3 days prior to admission to 30 days post discharge. Costs outside of the 30 day window may also be captured in this measure. For example, if a patient is admitted to an IRF following discharge and that stay is longer than 30 days, all costs from that stay are included. Therefore, if the patient is in the IRF for 45 days, the developer includes 45 days of charges and not 30. We have concerns that this was not addressed by the measure developer in the worksheet and request that additional testing be done to determine the validity and appropriateness of including costs beyond the 30 day window.

** Yes. Uses validity testing at measure score. Uses empirical validity testing. Provides three new analyses to support validity.

**Moderate--problematic that most of the variation is in the SNF setting/component and hospital has less "control" in that space. I don't believe Medicare allows hospitals to steer patients to SNFs such that they could direct patients to higher quality SNFs.

***2b3. Exclusions Analysis***

Comments:

**The exclusions seem reasonable to me. I wonder about excluding deaths. The literature is rife with reports of the amount of resources expended during the last days of some patients' lives. What is the rationale for excluding those episodes? The Developer's assessment would seem to show that this is not a significant impact, though.

** Episodes are excluded if beneficiary does not have 90 days of enrollment prior to the triggering event. The rationale is that this 90 day period provides sufficient information for the risk adjustment tool. It's not clear to me that 90 days is a long enough period to capture HCC diagnoses. Transfers are excluded --- probably best given the controversy over where they would attribute. Death episode exclusions may require further discussion by the committee.

** Exclusions seem appropriate.

** The exclusion of "cancer" hospitals could bias the results of general hospitals treating sizable numbers of similar cancer patients. The expected values for cancer MSDRGs could appear too low if "cancer" hospitals have high costs.

** The statistical justification of the exclusions is satisfactory.

** The MSPB measure attribution methodology excludes acute-to-acute transfers and ESRD patients (but not other complex patient populations) from the measure methodology. The AAMC requests additional information and a rationale for why these exclusions are included, along with an explanation as to why other high risk patient populations are not excluded from the measure.

** Exclusions are acceptable.

**Moderate.  Exclusion of hospital transfers--is there any evidence of gaming (to avoid counting those cases)?  Why not hold both hospitals accountable so episode is included?

***2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures***
Comments:
**The risk adjustment strategy described seems standard.
** The sociodemographic analysis will require further discussion.
** CMS has used zip code level measures of income in its risk adjustment model.  There was extensive prior discussion of the limitations of this and recommendations to use beneficiary address information to use census tract information for this adjustment.  This was not addressed in this submission.
** Episode case mix/severity (MSDRG) and clinical comorbidity (CMS-HCC) are both incorporated in the model, as are some other important program factors (disability, ESRD and long-term care).
Although adjustment for the SDS measures have small effects on the majority of hospitals, are there large effects on the most extreme values, implying possible evidence of disparities?
** This measure uses a widely used and better understood risk adjustment model (CMS-HCC) that is already used in multiple other areas by CMS.
**Moderate--problem noted with disability indicator (data source) which is fixable).    Unclear whether rural entities are harmed by this measure as they have no options for redirecting care to higher quality providers (only game in town).
** The AAMC is very concerned with the conclusions drawn from the SDS analysis for the MSPB measure. The developers looked at two variables (Income-to-Poverty Ratio and Race), and concluded that inclusion of the Income-to-Poverty Ratio variable resulted in a plus or minus difference of 0.01 or less for 97% of hospitals. Due to the tight clustering of performance scores for this measure, a difference of 0.01 (or greater) may significantly affect hospitals caring for disadvantaged patient populations that may require higher utilization of services. The AAMC has concerns that utilization measures may be influenced by the patient population served. We urge the measure developer to relook at the results of the SDS adjustment to see whether certain categories of hospitals are disproportionately affected by these factors. As a first step, the full list of hospitals and their corresponding change in performance with inclusion of the SDS variables should be released for stakeholder review.

In addition to SDS, the AAMC requests that the measure developer review the impact of concentrated healthcare services within a geographic area on medical service utilization. For example, a greater number of hospitals and clinics within a set geographic area may lead to more care for patients and higher MSPB scores. Greater utilization of these services may lead to high MSPB scores but also improved patient clinical outcomes.
** Yes, well developed risk-adjustment model. Adjustment is based on CMS-HCC risk adjustment methodology. Analysis of SES and SDS is adequately covered.

***2b5. Identification of Statistically Significant & Meaningful Differences In Performance***
Comments:
** I see that this measure may identify significant differences in total cost for a given episode, but how does that impact Quality?  Not sure where that is reviewed in this submission.
** This measure does not address quality; however, it does reveal meaningful differences in resource use across episodes.
** The interquartile range is approximately a $2000 difference on a measure with a mean of $20,000, or 10%.  Would like more information about the sources of these differences between high and low performing hospitals to better understand/interpret differences.
** Do differences between teaching status and rural/urban classifications indicate that risk adjustment is inadequate and may need SDS and SES variables included to result in detecting meaningful differences?
** This measure does identify meaningful differences in Medicare spending, but since clinical quality is not defined, I do not think it allows meaningful differences in quality to be determined.
** Yes, this measure identifies meaningful differences. Clinically meaningful differences was determined by stratifying MSPB-Hospital measure scores by meaningful hospital characteristics, and comparing those results to expected findings discussed in the literature.
**Moderate--the measure developers indicate there are differences by the strata used, but are these the right stratification variables for examining meaningful differences?

**N/A
***2b7. Missing Data Analysis and Minimizing Bias***
Comments:

** Minimal impact
** No. But measure excludes Medicare Advantage patients and part D drug costs.
** Based on the evidence provided, the impact of missing data would appear to be of limited.
**High-no problems

**Criterion 3. Feasibility**
**Maintenance measures – no change in emphasis – implementation issues may be more prominent**

**3. Feasibility:** This requirement involves demonstrating:

- the extent to which the specification, including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
- the required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.
- the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

**Summary of Feasibility**

- This measure is based on administrative claims data, which the developer notes are "readily available and retrievable without burden".
- The developer indicates that all data elements are in defined fields in electronic claims.
- The developer states the measure's risk adjustment model utilizes the new version of the CMS-HCC methodology, which accounts for the conversion to ICD-10 codes.
- The measure is already in operational use. During 30-day preview periods, neither the developer nor CMS received reports about measure errors from the measured hospitals (i.e., IPPS hospitals with at least 25 episodes in the performance period).
- No feasibility concerns were raised by the Cost and Resource Use Steering Committee during the NQF Measure Endorsement review in 2013.

***Questions for the Committee:***

o *Are the required data elements routinely generated and used during care delivery?*

**Preliminary rating for feasibility:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

**Committee pre-evaluation comments**
**Criteria 3: Feasibility**

***3.Feasibility***

Comments:
**These data elements are routinely generated and available to the Developer.
**Measure is feasible.
**Routine claims data is relied upon.
**This measure appears to be feasible for entities like CMS, but independent calculation of this measure will be challenging for uses without immediate access to the CMS data needed to calculate and trend this measure internally.
**Uses claims data, so generally very feasible to implement.
**No concerns.
**High-very feasible to implement

**Criterion 4: Usability and Use**
**Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences**

**4. Usability and Use**: This requirement involves describing the extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

**Current uses of the measure** [from OPUS]
**Publicly reported?** ☒ **Yes** ☐ **No**

**Current use in an accountability program?** ☒ **Yes** ☐ **No** ☐ **UNCLEAR**
**OR**
**Planned use in an accountability program?** ☐ **Yes** ☐ **No**

**Accountability program details**

- The developer states the measure is currently used in 3 quality reporting programs:
  - Hospital Inpatient Quality Reporting (IQR) Program: This program pays hospitals that successfully report designated quality measures a higher annual update to their payment rates. For the January 1, 2015 to December 31, 2015 period of performance, 97.9% of eligible hospitals received an MSPB-Hospital measure.
  - Hospital Compare: This program provides the public with information on the quality of care at Medicare-certified hospitals. The MSPB-Hospital measure is reported on the Hospital Compare website. The number of reporting hospitals is the same as the IQR Program.
  - Hospital Value-Based Purchasing (HVBP) Program: This program provides financial incentives to eligible hospitals based on their performance on selected quality measures. The MSPB-Hospital measure is currently within the Efficiency domain of the program. In FY2016, 3,036 hospitals received the MSPB-Hospital measure out of 3,041 (99.8%).

**Improvement results**

- The developer cites the data under the Opportunity for Improvement data and states the data demonstrate improvement in results given nearly half of all hospitals improved their MSPB-Hospital measure score.

**Unexpected findings (positive or negative) during implementation**

- The developer notes there are no unexpected findings during implementation.

**Potential harms**

- The developer notes that there were no unintended consequences during development or implementation.

**Vetting of the measure by those being measured**

- Hospitals have an opportunity to report measure calculation errors during 30-day review periods that occur after receiving their scores.

**Measure can be deconstructed to facilitate transparency and understanding** ☒ **Yes** ☐ **No**

**Feedback:**

- During the 2013 review, the Cost and Resource Use Standing Committee identified the following concerns with respect to the measure's usability:
  - Many hospitals may not have the analytic capacity to understand the data and understand the impact of care outside of the hospitalization on the measure result.
  - The small variation in performance makes it difficult for the consumer to distinguish best performers.

- The developer responded that a hospital's measure information is provided in a variety of formats and comparator data (i.e., state and national level data) are provided to facilitate a hospital's ability to understand their own data and compare their performance against other hospitals. For consumers, downloadable files containing more performance information are available online for consumers.
- During the 2012-2013 MAP review, MAP supported this measure for inclusion in the IQR and HVBP programs. MAP did not support the inclusion of the measure in the PPS-Exempt Cancer Hospital Quality Reporting (PCHQR) or Long-Term Care Hospital Quality Reporting (LTCHQR) Programs citing that the measure, as specified, excluded important groups of patients served by PPS-Exempt hospitals and long-term care hospitals.

***Questions for the Committee***:

o *How can the performance results be used to further the goal of high-quality, efficient healthcare?*

o *Do the benefits of the measure outweigh any potential unintended consequences?*

o *How has the measure been vetted in real-world settings by those being measure or others?*

**Preliminary rating for usability and use:** ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

## Committee pre-evaluation comments

**Criteria 4: Usability and Use**

***4.Usability and Use***

Comments:

**The results appear to be pretty tightly clustered. Difference from 25th to 75th percentile is 0.09. How will this be reported in a meaningful fashion? It seems small differences could easily drive a change in grade/quartile. How will the public be apprised as to those issues?

**Overall the information is useful but the measure does not provide enough information to succeed in episode based payment programs.

**Measure is usable, as demonstrated by its ongoing use.

**MSPB can be used to identify high cost patterns. Enhanced reporting subdividing the results by MDC could be useful to hospitals and consumers.

**While a Medicare spending per beneficiary measure provides useful high level information, it does not inform the user as to the source(s) of the excess spending.

**While the MSPB measure is specified and tested at the facility level, it is currently used in the physician value modifier program. Under the measure, physicians are attributed certain costs that may be outside of their direct control. Until the measure is specified and tested at the clinician level, the AAMC recommends that NQF clearly state that use of this measure is inappropriate for physician reporting and performance programs.

The AAMC also has concerns with the directionality of this measure and requests that CMS and the measure developer address whether lower MSPB scores translates into better clinical quality outcomes. The Hospital Compare Star Ratings TEP cited these concerns in their Public Comment Report #1: Measure Selection for Hospital Star Ratings. The TEP ultimately chose not to include the MSPB in the star ratings noting that the measure "seek(s) to reduce variation by evaluating outcomes for which performance is "non-directional," meaning that a higher or lower score is not necessarily better." In order to make the MSPB measure meaningful in the VBP program, the cost measure should be directly paired with a clinical quality measure to help stakeholders determine whether lower utilization leads to improved outcomes.

**The measure is used in a few accountability programs. Has been vetted through the MAP.

**High/moderate. High for payer and provider. Low for consumer--not useful for that audience.

**Evidence provided does not appear to indicate any unintended consequences. That being said, there are still concerns that I have which were expressed in the previous steering committee meeting:

1. Many hospitals still may not have the analytic capacity to understand the data and understand the impact of care outside of the hospitalization on the measure result.
2. The small variation in performance makes it difficult for the consumer to distinguish best performers.

**Criterion 5: Related and Competing Measures**

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

**Related or competing measures**

- The developer states there are no other NQF-endorsed measures addressing the same measure focus in this same target population (i.e., Medicare beneficiaries enrolled in Medicare Part A and Part B who are discharged from short-term acute hospitals).

**5.a. Harmonization**: This requirement involves demonstrating that the measure specifications are harmonized with related measures OR the differences in specifications are justified.

- N/A

**Endorsement + Designation**

**The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.**

**This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it:** is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

**Eligible for Endorsement + designation**: ☒ **Yes** ☐ **No**

**RATIONALE IF NOT ELIGIBLE**: N/A

## Pre-meeting public and member comments

- Ms. Jayne H. Chambers from Federation of American Hospitals comment on 2/22/17:

The Federation of American Hospitals ("FAH") requests that the Admissions and Readmissions Standing Committee provide input on what new factor(s) and/or new analyses might be needed on measure #2158, "Medicare Spending per Beneficiary (MSBP) - Hospital" in light of the recent report released by the Office of the Assistant Secretary for Planning and Evaluation (ASPE). Specifically, the ASPE report provided further confirmation that sociodemographic factors are strongly linked to hospital performance on resource use. Plus, NQF and this committee should address the potential unintended consequences of continuing to endorse measures without sufficient adjustment. The FAH encourages the committee to request additional analyses from the developer if needed. FAH believes it is critical that the NQF evaluations of measures such as this one continue to factor in new information and recommendations given the constantly evolving nature and understanding of the role of SES.

- Ms. Koryn Y. Rubin from American Medical Association comment on 3/6/17:

The AMA continues to remain concerned with the use of this measure. The measure is currently in use within physician programs, but testing has only been performed at the hospital level, which is a serious concern since it cannot be assumed that this measure will have the same impact in a physician practice as in the hospital. We remain concerned over the variation in discharge costs and how much control a hospital has over them. Some hospitals may have a direct connection with a rehab facility and therefore would have some control over the costs associated with rehab. In other instances, a hospital may have no connection or ownership over a rehab facility and based on the availability of space with the non-connected rehab facility and therefore, no true control over the costs associated with rehab or continued relationship.

**Reliability Testing:** The developer states that data element reliability was completed based on CMS' audit process, but no data is provided to support whether the audit actually yields reliable results. Therefore, we question whether the developer is justified in stating that reliability was completed without any results. We also question the reliability score of 0.4 with 25 episodes. Acumen's previous submission mentioned that they provided the confidence intervals (CIs) but they are not in this current submission and understanding how wide the intervals are would be incredibly helpful with understanding the reliability of the measure. Therefore, we urge Acumen to release this information for the committee and the public to review. During the last review, Acumen stated that if they increased the minimum number of episodes to 50 the number of hospitals included goes from 99% to 95.9%, but did not provide the reliability score with 50 episodes. The AMA believes it is better to have a higher reliability score than capturing the maximum number of hospitals. The low reliability score of 0.4 leads to too much noise with the measure and inaccurate and faulty conclusions about care.

We continue to remain concerned with the data provided for the test/re-test results and the validity of the measure. The test/retest results showed approximately 30% of hospitals in the lowest spending quintile in one sample were not in the lowest spending quintile in the next sample. In addition, 30% of hospitals in the highest spending quintile in one sample were not in the highest spending quintile in the next sample.

**Validity Testing:** We request further review of the validity testing results from the 2012 submission. In S.11, Interpretation of Scores, it is stated that the measure should not be used alone since the results alone do not necessarily reflect the quality of care provided. Yet, when they tested the correlation of MSPB to the readmission measure (also used in physician programs) last time, CMS found a very weak association between the two.  However, Acumen did not do any further testing on the correlation of cost with quality during this current review and given the omission of information it calls into question the usability and validity of the measure.

## Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

**IM1. High Priority**

**IM1.1. Demonstrated High Priority Aspect of Healthcare**

Affects large numbers

High resource use

**IM1.2. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in IM.1.3.**NQF's Measure Application Partnership (MAP) has already determined the MSPB-Hospital measure is an important measure that has potential for high impact.  A 2012 NQF Pre-rulemaking report stated that "MAP strongly supports the direction of this measure pending additional specification and testing."[1]  The content below contains further evidence of the high impact nature of this measure.

The growth of Medicare expenditures has put enormous strain on federal and state budgets, employers, and families.  Total Medicare expenditures in 2015 were $647.6 billion, which constituted 3.6 percent of GDP.  Estimates state that Medicare expenditures could grow up to 6.0 to 9.1 percent of GDP by 2090, indicating a need to address the current level of Medicare spending.  Of the total Medicare expenditures, $188.3 billion, or 30 percent, was spent on hospital benefits under Medicare Parts A and B.[2]  The MSPB-Hospital measure focuses on quantifying spending during and related to hospital stays to allow hospitals to identify areas where spending is most concentrated and coordinate with other healthcare providers, which can help counteract these rising costs. Despite the fact that the U.S. leads the world in health expenditures per capita, the value that patients receive for these expenditures may be below that of other countries.[3]  In particular, one source of inefficiency that creates rising healthcare costs includes payment systems that reward medical inputs rather than outcomes.[4]  Transforming Medicare and other public and private insurers from systems that reward volume of service to ones that reward efficient, effective care and reduce delivery system fragmentation offers the possibility of reducing cost and improving patient outcomes.

To advance this transformation, CMS instituted the MSPB-Hospital measure. Section 1886(o)(2)(B)(ii) of the Social Security Act, as established by Section 3001 of the Patient Protection and Affordable Care Act (ACA), requires that CMS implement a measure of Medicare spending per beneficiary as part of its Hospital Value-Based Purchasing (VBP) initiatives. By measuring the cost of care through a measure of Medicare spending per beneficiary, CMS aims to recognize hospitals that can provide high quality care at a lower cost to Medicare.

The MSPB-Hospital measure aims to incentivize hospitals to coordinate care and reduce unnecessary utilization during the period immediately prior to, during, and in the 30 days after a hospital admission. Currently, Medicare's prospective payment system (PPS) reimburses hospitals on a case mix-adjusted, flat-rate basis, incentivizing hospitals to serve patients as efficiently as possible. However, hospitals could also have an incentive to discharge patients early to reduce the cost to their facility. Such early discharge of patients may decrease quality of care and increases costs to Medicare. For example, a 2014 study showed that the cost of an additional day of an inpatient stay was offset by expected cost savings from readmission of 15 to 65 percent.[5] In addition, improved care coordination between acute and post-acute providers could stem the rising cost of post-acute care through avenues such as reducing unnecessary hospital readmission. In 2015, skilled nursing facility and home health costs accounted for $47.5 billion of Medicare's expenditures. [2]

Unlike other resource use measures reported on Hospital Compare, the MSPB-Hospital measure is not condition-specific. Because a hospital's MSPB-Hospital measure uses all Medicare Part A and Part B claims for episodes during the period of performance, the MSPB-Hospital measure evaluates hospitals' efficiency across admissions for all conditions. However, as it is currently used in conjunction with existing quality measures available on Hospital Compare and reported as part of the CMS Hospital Inpatient Quality Reporting (IQR) and Hospital VBP Programs, the MSPB-Hospital measure can identify efficient providers that provide high-quality, low-cost care.[6] Assessing the MSPB-Hospital measure alongside existing quality measures follows the NQF precedent of defining efficient care to be a measure of cost of care associated with a specified level of quality of care.

For the January 1, 2015 to December 31, 2015 period of performance, the MSPB-Hospital measure will be calculated from the claims of 4,261,069 Medicare beneficiaries and will affect 3,298 IPPS hospitals.

**IM1.3. Citations for data demonstrating high priority provided in IM.1.2**

- [1] National Quality Forum Measure Application Partnership. Pre-Rulemaking Report: Input on Measures Under Consideration by HHS for 2012 Rulemaking. Final Report. February 2012. http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=69885
- [2] Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2016 Annual Report. June 22, 2016. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ReportsTrustFunds/downloads/tr2016.pdf.

- [3] National Quality Forum. "Resource Use Measurement White Paper."
- [4] Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group, National Health Care Expenditures Data, August 2011, http://www.cms.gov/nationalhealthexpenddata/01_overview.asp.
- [5] Carey, Kathleen. "Measuring the Hospital Length of Stay/Readmission Cost Trade-Off Under a Bundled Payment Mechanism." Health Economics, Vol. 24, Issue 7 (July, 2015), pp. 790-802.
- [6] U.S. Department of Health & Human Services. Hospital Compare. www.hospitalcompare.hhs.gov.

IM2. **Opportunity for Improvement**

IM2.1. **Briefly explain the rationale for this measure (e.g., the benefits or improvements in performance envisioned by use of this measure)**

CMS includes the MSPB-Hospital measure within the Hospital VBP program as a measure of efficiency; the Hospital VBP program, however, also provides financial incentives to hospitals based on their performance on additional quality measures. By measuring the cost of care through the MSPB-Hospital measure in combination with these other quality measures, CMS aims to recognize hospitals that can provide high quality care at a lower cost to Medicare.

The MSPB-Hospital measure is designed to promote higher quality care for beneficiaries by financially incentivizing hospitals to improve care coordination, deliver efficient, effective care, and reduce delivery system fragmentation. Specifically, the MSPB-Hospital measure is calculated as the MSPB-Hospital amount compared to the national episode-weighted median MSPB-Hospital amount. This allows hospitals to improve their score by spending relatively less than the episode-weighted median during a given performance period. For instance, hospitals can decrease (i.e., improve) their MSPB-Hospital Amount through actions such as: 1) improving coordination with post-acute providers to reduce the likelihood of hospital readmissions, 2) identifying unnecessary or low-value post-acute services and reduce or eliminate these services, or 3) shifting post-acute care from more expensive services (e.g., skilled nursing facilities) to less expensive services (e.g., home health) in cases that would not affect patient outcomes.

Care coordination helps ensure a patient's needs and preferences for care are understood, and that those needs and preferences are shared between providers, patients, and families as a patient moves from one healthcare setting to another. People with chronic conditions, such as diabetes and hypertension, often receive care in multiple settings from numerous providers. As a result, care coordination among different providers is required to avoid waste, over-, under-, or misuse of prescribed medications and conflicting plans of care.

IM2.2. **Provide performance scores on the measure as specified** (current and over time) **at the specified level of analysis.** (This is required for endorsement maintenance. Include mean, stddev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include). **This information also will be used to address the subcriterion on improvement (U.2.1.) under Usability and Use.**
Analysis of all IPPS eligible hospitals with at least 25 episodes for the performance period of January 1, 2015 to December 31, 2015 shows a large range of provider scores on the MSPB-Hospital measure. The mean MSPB-Hospital measure score is 0.99, with a standard deviation of ±0.09. Since the MSPB-Hospital measure is calculated using the episode-weighted median as the denominator, the mean MSPB-Hospital measure score will not necessarily be 1.00. Provider scores range from a minimum of 0.59 to a maximum of 2.25, with a median value of 0.99. The 25th and 75th percentile of the measure score are 0.94 and 1.03, respectively, resulting in an interquartile range of 0.09.
Score distributions by decile are as follows:
10th – 0.89; 20th – 0.92; 30th – 0.95; 40th – 0.97; 50th – 0.99; 60th – 1.00; 70th – 1.02; 80th – 1.04; 90th – 1.08.

Analysis on the MSPB-Hospital amount showed similar results. The mean MSPB-Hospital amount is $20,168, with a standard deviation of $1,833. Provider amounts range from a minimum of $12,072 to a maximum of $46,074, with a median of $20,221. The 25th and 75th percentile of the MSPB-Hospital amounts are $19,195 and $21,136, respectively, resulting in an interquartile range of $1,941.
Score distributions by decile are as follows:
10th – $18,125; 20th – $18,886; 30th – $19,446; 40th – $19,839; 50th – $20,221; 60th – $20,546; 70th – $20,920; 80th – $21,344; 90th – $22,054.

Analysis of MSPB-Hospital provider score changes between 2014 and 2015 showed that hospital scores do vary over time. From 2014 to 2015, 47.46% of hospitals improved on their MSPB-Hospital measure score, which is defined as having a lower score in 2015 than in 2014. The minimum percent change (i.e., improvement) was -49.46%, while the maximum percent change was 264.29%.
Percent changes by decile are as follows:
10th – -4.08%; 20th – -2.30%; 30th – -1.25%; 40th – -0.47%; 50th – 0.16%; 60th – 0.84%; 70th – 1.66%; 80th – 2.83%; 90th – 4.99%.
Negative percent changes mean that the hospital improved on their MSPB-Hospital measure.

IM2.3. **If no or limited performance data on the measure as specified is reported in IM.2.2., then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
The response to IM2.2 includes measure scores calculated for all IPPS-eligible hospitals with at least 25 episodes during the performance period of January 1, 2015 to December 31, 2015.

IM2.4. **Provide disparities data from the measure as specified** (current and over time) **by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** (This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) **This information also will be used to address the subcriterion on improvement (U.2.1.) under Usability and Use.**
To analyze disparities by population group, we analyzed both socioeconomic status (SES) and sociodemographic status (SDS), where SDS is defined as SES and race considered together. To determine SES, we used American Community Survey (ACS) 5-year estimates to produce a distribution of household income to poverty level ratios (or "income-to-poverty ratio") for each 5-digit zip code. We then used beneficiary data (namely, 5-digit zip code and race) to estimate the income-to-poverty ratio for each beneficiary with an MSPB-Hospital episode. We defined race as either Black or Non-Black using data from the Enrollment Database (EDB). While the EDB provides data on all race categories, there are concerns with the validity of the other race categories (e.g., Asian, Hispanic) due to underreporting in those categories.[1] As a result, we categorized beneficiaries as either Black or Non-Black, where Non-Black is defined as all other race categories.

Using these data, we conducted analyses related to disparities by population group. For each race category (Black or Non-Black), we produced an estimated distribution of beneficiaries by income-to-poverty ratio. Among the lower income-to-poverty ratio ranges (i.e., below or near the poverty level), there was a greater percentage of beneficiaries who were Black (19%) when compared to Non-Black (11%). Among higher income-to-poverty ratio ranges (i.e., ratio above 5), there was a greater percentage of beneficiaries who were Non-Black (31%) compared to Black (22%).

Additionally, we sought to determine the effect of incorporating SES or SDS into our risk adjustment model by determining the difference in MSPB-Hospital measure scores when including SES or SDS. In both cases, the differences in MSPB-Hospital measure scores were minimal. When including SES in risk adjustment, the MSPB-Hospital measure score for 97% of hospitals changed by ±0.01 or less. When including SDS in risk adjustment, the MSPB-Hospital measure score for 95% of hospitals changed by ±0.01 or less.

[1] Zaslavsky, Alan M, John Z Ayanian, and Lawrence B Zaborski. “The Validity of Race and Ethnicity in Enrollment Data for Medicare Beneficiaries.” Health Services Research 47.3 Pt 2 (2012): 1300–1321. PMC. Web. 28 Oct. 2016.

IM2.5. **If no or limited data on disparities from the measure as specified is reported in IM.2.4., then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

IM3. **Measure Intent**

IM3.1. **Describe intent of the measure and its components/ Rationale (including any citations) for analyzing variation in resource use in this way.**

The MSPB-Hospital measure aims to incentivize hospitals to coordinate care and reduce unnecessary utilization during the period immediately prior to, during, and in the 30 days after a hospital discharge. As mentioned in IM1.2, because a hospital’s MSPB-Hospital measure is based on all Medicare Part A and Part B claims data for episodes during the period of performance and is not condition-specific, the MSPB-Hospital measure evaluates hospitals’ efficiency across all conditions and admissions. The all-cause nature of the MSPB-Hospital measure allows it to be applicable to a larger number of hospitals, maximizing its impact. The effect of patient health status and demographics on episode spending is accounted for by the MSPB-Hospital’s risk-adjustment methodology. One can measure whether hospitals provide efficient care by examining the MSPB-Hospital measure alone as well as in concert with a variety of quality of care measures already reported on CMS’ Hospital Compare webpage and developed as part of CMS’s Hospital Inpatient Quality Reporting and Hospital VBP Programs.

## Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

**Specifications** The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Non-Condition Specific** *(check all the areas that apply):*

Care Coordination<br>
Safety : Overuse

De.7. **Care Setting** *(Select all the settings for which the measure is specified and tested):*

Hospital : Acute Care Facility

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
<WebPageURLExists
nodeType="1">http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1228772057350

S.2. **Type of resource use measure** *(Select the most relevant)*

S.3. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED):*
Facility

S.4. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*

S.5. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
*If other, please describe in S.5.1.*
Claims (Only)
Other

S.5.1. **Data Source or Collection Instrument** *(Identify the specific data source or data collection instrument, e.g. name of database, clinical registry, collection instrument, etc.)*
The MSPB-Hospital measure uses Medicare Part A and Part B claims data, which is maintained by CMS' Office of Information System (OIS). Data from the Medicare Enrollment Database (EDB) are used to predict costs of episodes and determine beneficiary-level exclusions, specifically to determine the following: Medicare Parts A, B, and C enrollment; primary payer; disability status; end-stage renal disease (ESRD); beneficiary birth dates; and beneficiary death dates. The risk adjustment model also accounts for expected differences in payment for services provided to beneficiaries in long term care, and that information comes from the Minimum Data Set (MDS). The MDS is used to create the Long Term Care Indicator variable in risk adjustment (denoted as LTC_Indicator).
Data from the American Community Survey (ACS) is in the analyses performed to evaluate including SES/SDS in risk adjustment (see Testing Attachment Section 2b4).

S.5.2. **Data Source or Collection Instrument Reference** *(available at measure-specific Web page URL identified in S.1 OR in the file attached here) (Save file as: S_5_2_DataSourceReference)*
S_5_2_DataSourceReference-636149872134560000.pdf

S.6. **Data Dictionary or Code Table** *(Please provide a web page URL or attachment if exceeds 2 pages. NQF strongly prefers URLs. Attach documents only if they are not available on a web page.)*

***Data Dictionary:***

URL: The MSPB-Hospital measure relies on Medicare claims data. The Research Data Assistance Center (ResDAC) maintains an updated Medicare claims data dictionary available at the following URL: http://www.resdac.org/cms-data/file-family/Medicare-Claims.

Please supply the username and password:

Attachment:

***Code Table:***

URL:

Please supply the username and password:

Attachment:

**Construction Logic**

S.7.1. **Brief Description of Construction Logic**

If applicable, summarize the general approach or methodology to the measure construction. This is most relevant to measures that are part of or rely on the execution of a measure system or applies to multiple measures.

The MSPB-Hospital measure is the ratio of payment-standardized, risk-adjusted MSPB-Hospital amount for each hospital divided by the episode-weighted median MSPB-Hospital amount across all hospitals.

The numerator for a hospital's MSPB-Hospital measure is the hospital's MSPB-Hospital amount, which is the average spending level for the hospital's MSPB-Hospital episodes divided by the average expected episode spending level for the hospital's episodes, multiplied by the average spending over all episodes across all hospitals nationally. An MSPB-Hospital episode includes all Medicare Part A and Part B claims with a start date falling between 3 days prior to an Inpatient Prospective Payment System (IPPS) hospital admission (also known as the "index admission" for the episode) through 30 days post-hospital discharge.

The denominator for a hospital's MSPB-Hospital measure is the episode-weighted median MSPB-Hospital amount across all episodes nationally.

S.7.2. **Construction Logic** *(Detail logic steps used to cluster, group or assign claims beyond those associated with the measure's clinical logic.)*

The MSPB-Hospital measure is calculated according to the following eight steps:

Step 1: Standardize Claims Payments

To account for payment variation which is not directly related to decisions to utilize care, such as local or regional price differences or payments that reflect broader agency goals, standardized payments for each claim are calculated using the CMS payment standardization methodology to exclude geographic payment rate differences and certain add-on and incentive adjustments. In other words, the MSPB-Hospital measure adjusts observed payments for Medicare geographic adjustment factors, such as the hospital wage index and geographic practice cost index (GPCI) and payments such as disproportionate share (DSH) add-ons or Hospital Value-Based Purchasing (VBP) adjustments. More information about this is included in Section S.9.6.

Step 2: Calculate Standardized Episode Spending

Standardized spending during an episode is calculated as the sum of all the standardized Medicare claims payments (allowed amounts) made during the MSPB-Hospital episode (i.e., between 3 days prior to the hospital admission until 30 days after discharge). Standardized episode spending is also referred to as standardized episode cost.

Step 3: Calculate Expected Episode Spending

To estimate the relationship between standardized episode cost and a large set of independent variables (i.e., age, Hierarchical Condition Categories (HCCs), enrollment status, ESRD status, comorbidity interactions, long-term care, and MS-DRGs of the index admission), the MSPB-Hospital methodology uses an ordinary least squares (OLS) regression. Using a separate model for episodes within each major diagnostic category (MDC), standardized episode cost is regressed on these variables in a multivariate regression. The predicted values from this regression represent the expected spending for each episode.

Step 4: Winsorize Predicted Values

Although including a large number of variables in the regression more accurately captures beneficiary case mix, a large number of variables can also produce some extreme predicted values due to having only a few outlier episodes in a given cell. To prevent creating extreme predicted values, this step winsorizes (also known as 'bottom-codes') predicted values at the 0.5th percentile.[1],[2] This step also renormalizes the predicted values to ensure that the average expected episode spending level for each MDC is the same before and after winsorizing. This renormalization occurs by multiplying the winsorized predicted values by the ratio of the average standardized spending level within each MDC and the average bottom-coded predicted spending level within each MDC.

Step 5: Calculate Residuals

The residuals for each episode are calculated as the difference between the standardized episode spending level in Step 2 and the bottom-coded predicted value of spending for that episode calculated in Step 4. If the variable Y_ijm represents standardized spending levels for episode i for hospital j of MDC m, and Y(hat)_ijm equals the predicted spending levels from Step 4, then the residual is calculated as the following equation: Residual_ijm = Y_ijm – Y(hat)_ijm.

Step 6: Exclude Outliers

To mitigate the effect of high-cost and low-cost outliers on each hospital's MSPB-Hospital measure score, outliers are excluded at the episode level. Specifically, MSPB-Hospital episodes whose residuals fall above the 99th percentile or below the 1st percentile of the distribution of residuals across all MSPB-Hospital episodes are excluded from the MSPB-Hospital calculation. Excluding outliers based on residuals eliminates the episodes that deviate most from their predicted values in absolute terms. This step also renormalizes the predicted values to ensure that the average expected episode spending levels are the same as average standardized spending levels after outlier exclusions. This renormalization multiplies the predicted values after excluding outliers by the ratio of the average standardized spending level and the average bottom-coded predicted spending level after excluding outliers.

Step 7: Calculate the MSPB-Hospital amount for Each Hospital
The MSPB-Hospital amount for each hospital depends on three factors: i) the average standardized episode spending level from Step 2, ii) the average expected episode spending for each hospital calculated after Step 6, and iii) the average standardized episode spending across all hospitals. To calculate the MSPB-Hospital amount for each hospital, one finds the ratio of the average standardized episode spending over the average expected episode spending and then multiplies this ratio by the average episode spending level across all hospitals. Mathematically, the MSPB-Hospital amount is calculated as: MSPB amount_j = [(1/n_j)*(the sum of Y_ij over all elements i in the set {I_j})]/[(1/n_j)*(the sum of Y(hat)_ij over all elements i in the set {I_j})] * [(1/n)*(the sum of Y_ij over all i)], where Y_ij is the standardized spending for episode i in hospital j; Y(hat)_ij is the spending for episode i in hospital j, using the bottom-coded, renormalized predicted values from the risk adjustment regression after Step 6; n_j is the number of episodes for hospital j; n is the number of episodes across all hospitals in the U.S.; and i is an element of {I_j} indicates all episodes i in the set of episodes attributed to hospital j.
This equation defines the MSPB-Hospital amount for hospital j as the average spending level for hospital j divided by the average expected episode spending level for hospital j, multiplied by the average spending over all episodes across all hospitals. The MSPB-Hospital amount represents the per-episode spending level for a hospital assuming its composition of episodes matches that of the national average.

Step 8: Calculate the MSPB-Hospital measure
The MSPB-Hospital measure for hospital j is calculated as the ratio of the MSPB-Hospital amount for hospital j (calculated in Step 7) divided by the episode-weighted median MSPB-Hospital amount across all hospitals: MSPB Measure_j = MSPB Amount_j / National Median MSPB Amount. The national median MSPB-Hospital amount is a weighted median, where the weights are the number of episodes in each hospital.

[1] Winsorization is a statistical transformation that limits extreme values in data to reduce the effect of possibly spurious outliers. Winsorization typically involves both bottom-coding and top-coding, but the MSPB-Hospital measure uses only bottom-coding. Thus, all predicted values below the 0.5th percentile are assigned the value of the 0.5th percentile.
[2] To ensure that the lowest predicted values within an MDC are adjusted even for MDCs with few episodes, this methodology first sets the lowest predicted value within the MDC to the second lowest predicted value within the MDC before bottom-coding at the 0.5th percentile.

S.7.2a. **CONSTRUCTION LOGIC ATTACHMENT or URL:** If needed, attach supplemental documentation (Save file as: S_7_2_Construction_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.

URL:

Please supply the username and password:

Attachment:

S.7.3. **Concurrency of clinical events, measure redundancy or overlap, disease interactions** *(Detail the method used for identifying concurrent clinical events, how to manage them, and provide the rationale for this methodology.)*
We do not provide specifications for concurrency of clinical events.

The MSPB-Hospital measure methodology does not separate concurrent events and does not allow admissions within 30 days after discharge to start a new MSPB-Hospital episode.

The MSPB-Hospital measure methodology defines an MSPB-Hospital episode as all claims with start date falling between 3 days prior to an IPPS hospital admission (index admission) through 30 days post-hospital discharge. It includes the period 3 days prior-hospital admission and 30 days post-hospital discharge to emphasize the importance of care transitions and care coordination in improving patient care. Please refer to S.8.4., which details the rationale for the construction of the MSPB-Hospital episode, for a discussion of the advantages of this approach.

Note that the MSPB-Hospital measure calculation does not pro-rate the cost of care that extends beyond the 30 days post-hospital discharge. For example, if a patient is admitted to an IPPS hospital, triggers an MSPB-Hospital episode, and then receives Inpatient Rehabilitation Facility (IRF) care that begins within the 30 days after discharge, the index hospital is responsible for the full cost of the IRF claim even if the claim extends longer than 30 days after discharge. Pro-rating this cost of care could result in episodes at the end of the performance period having lower risk-adjusted episode costs since only a portion of costs of claims that occur across performance periods (e.g., a claim starts in 2015 and ends in 2016) would be counted into the observed costs of the episode.

S.7.4. **Complementary services** *(Detail how complementary services have been linked to the measure and provide rationale for this methodology.)*

An episode includes all services from the 3 days prior to a hospital admission to promote MSPB-Hospital episode consistency regardless of the diagnosis code on the pre-admission services and where these complementary services take place. This is in part because Medicare includes certain services in its payment to IPPS hospitals. Specifically, diagnostic services and non-diagnostic services related to the reason for admission are captured in the inpatient diagnosis-related group (DRG) payment for the hospitalization when they are performed by the hospital during the 3 days prior to admission. Diagnostic services or non-diagnostic services related to the reason for admission that are performed by a provider other than the hospital are not captured in the inpatient DRG payment and are paid separately under Medicare. Furthermore, non-diagnostic services that appear to be unrelated to the reason for admission are also not captured in the inpatient DRG payment and are paid separately under Medicare. The MSPB-Hospital episode includes all services from 3 days prior to ensure that all costs are included in the measure. For additional discussion, please refer to S.8.4., which details the rationale for the construction of the MSPB-Hospital episode.

S.7.5. **Clinical hierarchies** *(Detail the hierarchy of codes or condition groups used and provide rationale for this methodology.)*

Clinical hierarchies are embedded in the risk adjustment model; see S.9.5. for more details. Severity of illness is measured using 79 Hierarchical Condition Category (HCC) indicators derived from the beneficiary's claims during the period 90 days prior to the start of the episode, an indicator of whether the beneficiary recently required long-term care, and the MS-DRG of the index hospitalization. The MSPB-Hospital risk-adjustment methodology is discussed in additional detail in S.9.3. and S.9.4.

Episode construction does not utilize clinical hierarchies, as the MSPB-Hospital measure includes all services in the time window regardless of diagnosis or DRG, as described above in S.7.4.

S.7.6. **Missing Data** *(Detail steps associated with missing data and provide rationale for this methodology (e.g., any statistical techniques to impute missing data)*
We do not provide measure specifications or guidelines for missing data :

All the data used to calculate hospitals' MSPB-Hospital measure values are included on Medicare claims data. The data fields used to calculate the MSPB-Hospital measure (e.g., payment amounts, DRGs, diagnosis and procedure codes, etc.) are included in all Medicare claims because hospitals only receive payments for complete claims. Additional information regarding the reliability of diagnostic information on claims is available on the Testing Form in Section 2a2.2.

The data used to calculate the MSPB-Hospital measure includes all data for Medicare claims. We do have complete data for each beneficiary who has an index admission, since beneficiaries are excluded if they are not continuously enrolled in only Medicare Parts A and B or if Medicare is not the primary payer during an episode, as described in S.9.1. This ensures that we have all claims data for beneficiaries included in the MSPB-Hospital measure calculation.

S.7.7. **Resource Use Service Categories (Units) (Select all categories that apply)**

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

Durable Medical Equipment (DME)

S.7.8. **Identification of Resource Use Service Categories (Units)**
*(For each of the resource use service categories selected above, provide the rationale for their selection and detail the method or algorithms to identify resource units, including codes, logic and definitions.)*
The MSPB-Hospital measure assesses the standardized allowed amounts of services performed by hospitals and other healthcare providers during an MSPB-Hospital episode, which includes all Part A and Part B Medicare claims that occur within the time period 3 days prior to the index hospital admission through 30 days after discharge from the index admission. As a result, costs from all Part A and Part B claim types (i.e., inpatient, outpatient, home health agency, hospice, skilled nursing facility, durable medical equipment, and carrier) are included. Note that costs of Part B drugs are included but costs of Part D drugs are not included since Part D is not used to calculate the MSPB-Hospital measure. The methodology used to payment standardize these claims is available for download ("CMS Price (Payment) Standardization") from the URL provided in S.7.8a.

To assist providers in examining their spending, CMS provides MSPB-Hospital spending breakdowns by different claim types (i.e., home health agency, hospice, inpatient, outpatient, skilled nursing facility, durable medical equipment, and physician/carrier), as well as by time period (i.e., 3 days prior to index admission, during-index admission, and 30 days after hospital discharge). These data are provided at the following URLs:

- https://data.medicare.gov/Hospital-Compare/Medicare-Hospital-Spending-by-Claim/nrth-mfg3

S.7.8a. **If needed, provide supplemental resource use service category specifications in either URL (preferred) or as an attachment (Save file as S.7.8a_RU_Service_Categories):**
URL: http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1228772057350

Please supply the username and password:

Attachment:

**Clinical Logic**

S.8.1. **Brief Description of Clinical Logic** (Briefly describe your clinical logic approach including clinical topic area, whether or not your account for comorbid and interactions, clinical hierarchies, clinical severity levels and concurrency of clinical events.)
Objective: The MSPB-Hospital measure aims to improve care coordination and care quality in the period between 3 days prior to an acute inpatient hospital admission through the period 30 days after discharge.

Clinical Topic Area: Inpatient Admissions, all conditions

Accounting for Comorbidities: Application of a variant of the CMS-HCC risk adjustment model. The model includes a full set of interaction terms between comorbidities and MDC of the index admission, as well as a select number of interaction terms between comorbidities.

Measure of Episode Severity: Risk adjustment model includes indicators for the MS-DRG of the index admission.

Concurrency of Clinical Events. The MSPB-Hospital episode spans the period 3 days prior to the index hospital admission through 30 days post-discharge. All events that occur during this time period are included in the MSPB-Hospital episode.

S.8.2. **Clinical Logic** *(Detail any clustering and the assignment of codes, including the grouping methodology, the assignment algorithm, and relevant codes for these methodologies.)*
Objective: The MSPB-Hospital measure aims to improve care coordination in the period between 3 days prior to an acute inpatient hospital admission through the period 30 days after discharge. The MSPB-Hospital measure recognizes lower costs associated with a reduction in unnecessary services, preventable complications, readmissions, and shifting post-acute care from more expensive to less expensive services when appropriate.

Grouping methodology: The MSPB-Hospital measure evaluates resource use through the unit of MSPB-Hospital episodes. The MSPB-Hospital episodes are constructed by including all Medicare Part A and Part B claims with a start date falling between 3 days prior to an acute inpatient hospital admission through the period 30 days after discharge.

Any episodes where at any time during the episode the beneficiary is enrolled in a Medicare Advantage plan, the beneficiary becomes deceased, or Medicare is the secondary payer will be excluded from the MSPB-Hospital calculation. Regarding beneficiaries whose primary insurance becomes Medicaid during an episode due to exhaustion of Medicare Part A benefits, Medicaid payments made for services rendered to these beneficiaries are excluded; however, all Medicare Part A payments made before benefits are exhausted and all Medicare Part B payments made during the episode are included.

Cost Calculation: The MSPB-Hospital amount includes the cost of services performed by hospitals and other healthcare providers during an MSPB-Hospital episode, which is comprised of the period 3 days prior to an inpatient PPS hospital admission (index admission) through 30 days post-hospital discharge. All costs are payment standardized to control for geographic variation in Medicare reimbursement rates. All costs are risk adjusted to account for age and severity of illness. More details about the risk adjustment model is described in section S.9.3.

S.8.3. **Evidence to Support Clinical Logic Described in S.8.2** *Describe the rationale, citing evidence to support the grouping of clinical conditions in the measurement population(s) and the intent of the measure (as described in IM3)*
The MSPB-Hospital measure methodology defines an MSPB-Hospital episode as all claims with start dates falling between 3 days prior to an IPPS hospital admission (index admission) through 30 days post-hospital discharge and does not separate concurrent events. It includes the period 3 days prior-hospital admission and 30 days post-hospital discharge to emphasize the importance of care transitions and care coordination in improving patient care and reducing unnecessary readmissions. This episode definition is consistent with MedPAC's response to the FY 2012 IPPS proposed rule, in which they recommended that "both CMS and MedPAC should focus on creating parallel incentives for hospitals and post-acute care providers to work to reduce readmissions. The end goal is to align incentives across the sectors to encourage cooperation among providers to improve the quality of the episode of care, reduce the cost of the episode of care, and reduce the number of unnecessary inpatient episodes" (http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY-2012-IPPS-Final-Rule-Home-Page.html).

The advantage of this approach is that this approach is simple, as costs of Medicare services do not need to be divided into separate clinical events. Under the MSPB-Hospital measure methodology, costs do not need to be divided between those more relevant and those less relevant to the episode. In addition, the approach aligns with other measures (such as quality measures) based on Medicare claims billed during and after a hospital admission and is consistent with feedback received from stakeholders through notice and comment rulemaking.

S.8.3a. **CLINICAL LOGIC ATTACHMENT or URL: If needed, attach supplemental documentation (Save file as: S_8_3a_Clinical_Logic). All fields of the submission form that are supplemented within the attachment must include a summary of important information included in the attachment and its intended purpose, including any references to page numbers, tables, text, etc.**
URL:

Please supply the username and password:

Attachment: 2016_11_02_mspb_hospital_testing_appendix_tables.xlsx

S.8.4. **Measure Trigger and End mechanisms** *(Detail the measure's trigger and end mechanisms and provide rationale for this* *methodology)*

Trigger Event: An MSPB-Hospital episode, which serves as the unit of analysis for the MSPB-Hospital measure, will trigger with an IPPS hospital admission. Admissions that occur within 30 days of discharge from another index admission and admissions during which a beneficiary is transferred from one acute hospital to another are not considered to be index admissions. Hospitalizations that occur within the 30-day post discharge window of the index admission are included in the same episode the index admission opened. On the other hand, hospitalizations that begin more than 30 days after the beneficiary is discharged from a hospital trigger a new MSPB-Hospital episode as an index admission.

MSPB-Hospital Episode Start Date: 3 days prior to index inpatient admission

MSPB-Hospital Episode End Date: 30 days after discharge from the index hospital admission

An episode includes the 3 days prior to a hospital admission to promote MSPB-Hospital episode consistency regardless of the diagnosis code on the pre-admission services and where these complementary services take place. This is in part because Medicare includes certain services in its payment to IPPS hospitals. Specifically, diagnostic services and non-diagnostic services related to the reason for admission are captured in the inpatient diagnosis-related group (DRG) payment for the hospitalization when they are performed by the hospital during the 3 days prior to admission. Diagnostic services or non-diagnostic services related to the reason for admission that are performed by a provider other than the hospital are not captured in the inpatient DRG payment and are paid separately under Medicare. Furthermore, non-diagnostic services that appear to be unrelated to the reason for admission are also not captured in the inpatient DRG payment and are paid separately under Medicare. The MSPB-Hospital episode includes Medicare payments for all Part A and Part B services from 3 days prior to ensure that all costs are included in the measure. Furthermore, an episode includes the 30 days after a hospital discharge to emphasize the importance of care transitions and care coordination in improving patient care. Only discharges occurring at least 30 days before the end of the measurement period are counted as index admissions. Admissions that occur within 30 days of discharge from another index admission are not considered to be index admissions.Trigger Event: Inpatient admission, with the exception of acute-to-acute transfer cases

Start Date: 3 days prior to index inpatient admission

End Date: 30 days after discharge from the index hospital admission

As discussed in S.8.2., an MSPB episode is defined as all claims with start date falling between 3 days prior to an inpatient PPS hospital admission (index admission) through 30 days post hospital discharge. In other words, the MSPB Measure’s trigger is an inpatient PPS hospital admission, and the start is 3 days prior to an index admission, while the end is 30 days post hospital discharge. Admissions that occur within 30 days of discharge from another index admission and admissions during which a beneficiary is transferred from one acute hospital to another are not considered to be index admissions. Hospitalizations that occur within the 30-day post discharge window of the index admission are attributed to the index admissions. On the other hand, hospitalizations that begin more than 30 days after the beneficiary is discharged from a hospital trigger a new MSPB episode as an index admission.

Diagnostic services and non-diagnostic services related to the reason for admission are captured in the inpatient DRG payment for the hospitalization when they are performed by the hospital during the 3 days prior to admission (http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Three_Day_Payment_Window.html); however, if, during the 3 days prior to a hospital admission, a beneficiary receives diagnostic services from a provider other than the hospital or non-diagnostic services that appear on the claim to be unrelated to the reason for admission, those services are separately payable under Medicare. To promote MSPB episode consistency regardless of where these complementary services take place and to incorporate payments for services that may appear on the face of a claim to be unrelated to the original admission (as described in section S.8.2), a 3-day window prior to the index admission is included at the start of the MSPB episode. The MSPB time frame also includes services that take place during the time period 30 days post-hospital discharge in order to emphasize the importance of care transitions and care coordination in improving patient care. As a result, services whose claim start dates fall between 3 days prior to an index admission through 30 days post hospital discharge are attributed to that index admission.

The advantages of this measure trigger and end mechanism are twofold. First, this approach is simple and easily-implementable since it includes all claims during the MSPB episode. An alternative would be to create separate episodes for each type of hospital admission. Although episode-based approaches are attractive for a number of purposes, the MSPB aims to evaluate overall hospital efficiency level across all types of care and creating are over 700 types of hospitals admission episodes (i.e., there are over 700 MS-DRGs) is not practical. Second, the MSPB approach incorporates costs due to care complications unrelated to the original admission, encouraging hospital care coordination. For example, if a beneficiary is admitted for AMI but develops pneumonia due to poor care coordination, these costs will be captured in the episode generated by the initial AMI index admission.

S.8.5. **Clinical severity levels** *(Detail the method used for assigning severity level and provide rationale for this methodology)*

Clinical severity levels are embedded in the risk adjustment model, as described in S.9.2. through S.9.5.

S.8.6. **Comorbid and interactions** *(Detail the treatment of co-morbidities and disease interactions and provide rationale for this methodology.)*

Controlling for Comorbid Conditions and Interactions: The MSPB-Hospital measure accounts for comorbid conditions and interactions by broadly following the CMS- Hierarchical Condition Categories (HCCs) risk-adjustment methodology, which is derived from Medicare Part A and B claims and is used in the Medicare Advantage (MA) program. Diagnosis codes on claims that occur during the 90-day period prior to the start of an MSPB-Hospital episode are used to create HCC indicators. Episodes where the beneficiary is not enrolled in both Medicare Part A and Medicare Part B for the 90 days prior to the episode are excluded because information on comorbidities for these beneficiaries will be incomplete. When applying the CMS-HCC framework to the MSPB-Hospital measure, expected costs are determined by the risk adjustment model separately for each Major Diagnostic Category (MDC), which allows the effect of beneficiary health status and demographics on episode spending levels to vary by the MDC of the MSPB-Hospital index admission. The MSPB-Hospital measure accounts for comorbid interactions by incorporating a number of health status interactions as currently used within the CMS-HCC model. The model includes paired-condition interactions (e.g., chronic obstructive pulmonary disease (COPD) and congestive heart failure (CHF)) and interactions between conditions and disability status (e.g., disabled and cystic fibrosis). The full list of variables used in the risk adjustment model can be found in S.9.4.

The 90-day period prior to the start of an episode is used to measure the conditions which most directly impact beneficiaries' health status at the time of the hospital admission and to capture beneficiaries' comorbidities in the risk adjustment. Additionally, because the relationship between comorbidities' episode cost may be non-linear in some cases (i.e., beneficiaries may also have more than one disease during a hospitalization episode), the model also takes into account a limited set of interactions between HCCs and/or enrollment status variables. The MSPB-Hospital measure risk adjustment methodology includes only a limited set of interaction terms for two reasons. First, inclusion of too many interaction terms will over-fit the model. Second, the MSPB-Hospital measure risk-adjustment methodology broadly follows the established CMS-HCC risk-adjustment methodology, which uses similar interaction terms.

Concurrent Clinical Conditions: To simplify the clinical logic, all claims that begin during the period 3 days prior to the index admission through 30 days after discharge are included in a given MSPB-Hospital episode. See Section S.8.3. above for more details.

**Adjustments for Comparability**

S.9.1. **Inclusion and Exclusion Criteria** *Detail initial inclusion/exclusion criteria and data preparation steps (related to clinical exclusions, claim-line or other data quality, data validation, e.g. truncation or removal of low or high dollar claim, exclusion of ESRD patients)*

:

- Included Populations:

The beneficiary population eligible for the MSPB-Hospital measure calculation is made up of Medicare beneficiaries enrolled in Medicare Parts A and B who were discharged from short-term acute hospitals during the period of performance. Specifically, Medicare Part A and Medicare Part B claims from beneficiaries with an index admission within a subsection (d) hospital are included in the MSPB-Hospital episode if the beneficiary has been enrolled in Medicare Part A and Part B for the period 90 days prior to the start of an episode (i.e., 93 days prior to the date of the index admission) until the 30 days after discharge.[1] For example, if the period of performance is May 1, 2015 to December 31, 2015, hospitalizations must have a discharge date on or before December 1 to be eligible to be included as index admissions. Defining the population in this manner ensures that each beneficiary's claims record contains sufficient fee-for-service data both for measuring spending levels and for risk adjustment purposes.

Only claims for beneficiaries admitted to subsection (d) hospitals during the period of performance are included in the calculation of the MSPB-Hospital measure. Subsection (d) hospitals are hospitals in the 50 States and D.C. other than: psychiatric hospitals, rehabilitation hospitals, hospitals whose inpatients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer. An acute hospital is defined as those with provider variable's third position "0". The claims for inpatient admissions to subsection (d) hospitals are combined into "stays" by beneficiary, admission date, and provider.
[1] Claims reported by hospitals participating in the Medicare Acute Care Episode (ACE) Demonstration are also included in the MSPB measure calculation. In ACE Demonstration hospitals, physicians submit claims as usual, but ACE claims are categorized as "no pay." As a result, they show up in the standardized payment; consequently, ACE demonstration episodes are included in the MSPB measure.

- Excluded Populations:

Populations excluded from the MSPB-Hospital calculation are any episodes where at any time 90 days before or during the episode, the beneficiary is enrolled in a Medicare Advantage plan or Medicare is the secondary payer. Episodes where the beneficiary becomes deceased during the episode are also excluded. Regarding beneficiaries whose primary insurance becomes Medicaid during an episode due to exhaustion of Medicare Part A benefits, the beneficiaries themselves are not excluded. Rather, Medicaid payments made for services rendered to these beneficiaries are excluded, while all Medicare Part A payments made before benefits are exhausted and all Medicare Part B payments made during the episode are included. We believe this is the most appropriate method for addressing benefits exhaust episodes, because these beneficiaries represent high resource use cases that should be included in a hospital's measure. In addition, this removes the potential for hospitals to exhaust a beneficiary's Part A benefits to exclude high resource use episodes from their measure.

Further, any episode in which the index admission inpatient claim has a $0 actual payment or a $0 standardized payment is excluded. In addition, acute-to-acute transfers (where a transfer is defined based on the claim discharge code) are not considered index admissions. In other words, these cases do not generate new MSPB-Hospital episodes; neither the hospital which transfers a patient to another subsection (d) hospital, nor the receiving subsection (d) hospital will have an index admission or associated MSPB-Hospital episode attributed to them. This exclusion addresses stakeholder concerns that neither the admitting nor receiving hospital is fully able to coordinate care. Stakeholders stated that it was inappropriate to hold the initially-admitting hospital accountable for services rendered by the receiving hospital. In addition, stakeholders expressed concern with holding the receiving hospital accountable for any issues that arose as a result of the initially-admitting hospital's care and/or follow up care rendered near the beneficiary's home, where the receiving hospital may not be in an ideal place to coordinate that care.

Admissions to hospitals that Medicare does not reimburse through the IPPS system (e.g., cancer hospitals, critical access hospitals, hospitals in Maryland) are not considered index admissions and are therefore not eligible to begin an MSPB-Hospital episode. If an acute-to-acute hospital transfer or a hospitalization in a PPS-exempt hospital type happens during the 30-day window following an included index admission, however, it will be counted in the measure. This is because the MSPB-Hospital measure includes all claims and services that occur 30 days after discharge from the index hospital; an episode includes the 30 days after a hospital discharge to emphasize the importance of care transitions and care coordination in improving patient care.

The following lists details the exclusions made to all episodes of care for which full data are not available or for which Medicare spending by itself cannot reasonably be considered a signal of efficiency:

- [I] Any episodes without all observable claims or a complete episode window (i.e., episodes in which Medicare is the secondary payer, episodes in which the beneficiary is enrolled in a Medicare Advantage plan, episodes in which the beneficiary is enrolled only in Medicare Part A, episodes in which the beneficiary becomes deceased). Episodes in which the beneficiary is enrolled only in Medicare Part A, for example, are excluded because these beneficiaries may receive services not observed in the data. Similarly, episodes in which the beneficiary dies at any point during the episode. Episodes in which the patient dies are—by definition—truncated episodes and do not have a complete episode window. Episodes in which the patient dies were identified as an index hospitalization with death discharge code (STUS_CD "20" "41") or if a beneficiary's death was within an MSPB-Hospital episode. Including episodes without all observable claims or a complete episode window could potentially make hospitals seem efficient not due to any action of their own, but because the data are missing services that would be included in the MSPB-Hospital measure calculation.

Episodes where Medicare is the secondary payer: if a beneficiary was the primary payer any time during the MSPB-Hospital episode, the beneficiary was excluded (i.e., if bene_prmry_pyr_entlmt_strt_dt (start date of primary payer enrollment) bene_prmry_pyr_entlmt_end_dt (end date of primary payer enrollment) fell within the episode).

- [II] Regarding beneficiaries whose primary insurance becomes Medicaid during an episode due to exhaustion of Medicare Part A benefits, these beneficiaries are not excluded. Rather, Medicaid payments made for services rendered to these beneficiaries are excluded; all Medicare Part A payments made before benefits are exhausted and all Medicare Part B payments made during the episode are included.

The MSPB-Hospital measure is calculated using only Medicare Part A and Part B claims; as a result no Medicaid claims are included in the MSPB-Hospital measure calculation.

- [III] Any episode in which the index admission inpatient claim has a $0 actual payment or a $0 standardized payment is excluded. $0 inpatient admissions may represent errors in the data, or payment corrections rather than actual services rendered. Only when the Claim Payment amount (pmt_amt) for the IP stay is greater than 0 OR standard_allowed_amt is greater than 0 is the amount included in the MSPB-Hospital measure calculation.
- [IV] Due to the uncertainty surrounding attributing episodes to hospitals in cases where the patient was transferred between acute hospitals during the index admission, acute-to-acute transfers during the index admission (where a transfer is defined based on the claim discharge code) are not considered index admissions for the purposes of the MSPB-Hospital measure. In other words, these cases will not generate new MSPB-Hospital episodes; neither the hospital which transfers a patient to another short-term acute hospital, nor the receiving short-term acute hospital will have an index admission attributed to them. This exclusion avoids assigning responsibility to an MSPB-Hospital episode in a case where multiple hospitals treat the patient during the index admission.
- [V] Cancer hospitals, MD Hospitals (provider variable starting with “21”), emergency hospitals (provider variable last position “E” OR “F”), and veteran’s hospital (provider variable position “V”) are also excluded.
- [VI] In response to stakeholder comments, the FY 2012 IPPS Final Rule states that the MSPB-Hospital measure will “exclude statistical outliers from the calculation” (76 FR 51626: www.gpo.gov/fdsys/pkg/FR-2011-08-18/pdf/2011-19719.pdf). To mitigate the effect of high-cost outliers on each hospital’s MSPB-Hospital measure score, MSPB-Hospital episodes whose relative scores fall above the 99th percentile or below the 1st percentile of the distribution of residuals are excluded from the MSPB-Hospital calculation. Excluding outliers based on residuals eliminates the episodes that deviate most from their predicted values in absolute terms.

S.9.2. **Risk Adjustment Type** (Select type)

Statistical risk model

If other:

S.9.3. **Statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables.)*

To account for case-mix variation and other factors, the MSPB-Hospital risk adjustment methodology adjusts the MSPB-Hospital measure for age and severity of illness. The independent variables used in the risk adjustment model include beneficiary age, health status (as measured by hierarchical condition categories (HCCs)), disability-status, end-stage renal disease (ESRD) status, residence in a long-term care facility, and indicators for the MS-DRG of the index hospital admission. Severity of illness is measured using 79 HCC indicators derived from the beneficiary’s claims during the period 90 days prior to the start of the episode, an indicator of whether the beneficiary recently required long-term care, and the MS-DRG of the index hospitalization. The 79 HCC indicators are specified in Version 22 of the HCC model, and the HCC V22 model includes a mapping of ICD-9 diagnosis codes to CCs and ICD-10 diagnosis codes to CCs. As described above, episodes where the beneficiary is not enrolled in both Medicare Part A and Medicare Part B for the 90 days prior to the episode are excluded. This “look back period” captures beneficiaries’ comorbidities in the risk adjustment. The MSPB-Hospital risk adjustment methodology also includes status indicator variables for whether the beneficiary qualifies for Medicare through disability or age and ESRD. In addition, the model accounts for disease interactions by including interactions between HCCs and/or enrollment status variables that are included in the MA model. This is included because the presence of certain comorbidities increase costs in a greater way than predicted by the HCC indicators alone.[1] The MSPB-Hospital risk adjustment method does not control for the beneficiary’s sex and race.

The MSPB-Hospital risk adjustment approach uses an ordinary least squares (OLS) linear regression model and broadly follows the CMS-HCC risk adjustment methodology, which is derived from Medicare Part A and B claims and is used in the Medicare Advantage (MA) program.[2] Although the MA risk adjustment model includes 24 age/sex variables, the MSPB-Hospital methodology does not adjust for sex and only includes 12 age categorical variables. The OLS model is stratified based on the MDC of the index admission. The use of separate models by MDC permits the effect of risk factors on episode spending to vary based on the bodily system treated during the index admission. More precisely, this approach allows the coefficient on each risk adjuster to vary by MDC.

All variables are calculated using Medicare claims data during the period 90 days prior to the start of an episode. No risk adjustment factors are determined using information contemporaneous with the MSPB-Hospital episode to avoid circularity problems that would—by construction—cause the risk adjustment factors to be correlated with episode spending. For a detailed list of explanatory variables in the risk adjustment model, please see the response to Section S.9.4.

[1] Centers for Medicare and Medicaid Services. Medicare Managed Care Manual, Chapter 7 – Risk Adjustment, Section 70.2.7 – Disease and Disabled Interactions. 2014. https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/mc86c07.pdf
[2] Centers for Medicare and Medicaid Services, Office of the Actuary. "Announcement of Calendar Year (CY) 2014 Medicare Advantage Capitation Rates and Medicare Advantage and Part D Payment Policies and Final Call Letter." April 2013. https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/Announcement2014.pdf

S.9.4. **Detailed Risk Model Specifications** *available at measure-specific Web page URL identified in S.1 OR in attached data dictionary/code list Excel or csv file.*
Available at measure-specific web page URL identified in S.1

S.9.5. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets)*
While the measure results are not stratified, expected costs for episodes are determined by using a separate risk adjustment model for episodes within each MDC. MDCs are aggregations of Diagnosis Related Groups (MS-DRG), which CMS uses to classify acute inpatient admissions.
The MS-DRG/MDC crosswalk is available for order here:
http://solutions9.3m.com/wps/portal/!ut/p/c1/04_SB8K8xLLM9MSSzPy8xBz94NS8-NBg_Qj9KLP4IC8Py1BTI2MD9zAvFwMjYzMzCxNHd2OTACP9ggxHRQBm3gTM/

S.9.6. **Costing method**
Detail the costing method including the source of cost information, steps to capture, apply or estimate cost information, and provide rationale for this methodology.
Standardized pricing
As discussed in S.7.2., the MSPB-Hospital measure removes sources of variation which are not directly related to decisions to utilize care, such as local or regional price differences, to capture differences in beneficiary resource use that a hospital can influence through appropriate practices and care coordination. The MSPB-Hospital measure uses payment standardized allowed amounts posted on the CMS Integrated Data Repository (IDR) using the CMS methodology available at this MSPB-Hospital QualityNet webpage:
http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1228772057350. The documentation on this webpage lists the standardization methodology for each Medicare claims setting, as well as the applicable inputs for each setting. Allowed amounts include both Medicare trust fund payments and beneficiary deductible and coinsurance. The MSPB-Hospital measure uses allowed amounts to capture the cost of a service, without allowing episodes during which the beneficiary paid a higher portion to appear more expensive than episodes where, for example, a beneficiary has met his or her deductible and paid less.

Specifically, the payment (or price) standardization methodology:

- Eliminates adjustments made to national payment amounts to reflect differences in regional labor costs and practice expenses (measured by hospital wage indexes and geographic practice cost indexes);

- Substitutes a national amount in the case of services paid on the basis of state fee schedules;

- Eliminates payments to hospitals for larger program goals, including graduate indirect medical education (IME); serving a disproportionate population of poor and uninsured (i.e., disproportionate share payments (DSH)); and payments associated with incentive payment programs;

- Preserves differences that result from health care delivery choices such as the:
  - setting where the service is provided (e.g., physician office versus outpatient hospital);

- type of healthcare provider who provides the service (e.g., physician versus nurse practitioner);
- number of services provided in the same encounter; and
- outlier cases.

S.10. **Type of score***(Select the most relevant):*
Ratio
Attachment
If other:
Attachment: S10_sample_score_report-636136980406604000.pdf

S.11. **Interpretation of Score** *(Classifies interpretation of a ratio score(s) according to whether higher or lower resource use amounts is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score, etc.)*
An MSPB-Hospital measure that is less than 1 indicates that a given hospital's MSPB-Hospital amount (i.e. risk-adjusted spending) is less than the national episode-weighted median MSPB-Hospital amount across all hospitals during a given performance period. We note that results of the MSPB-Hospital measure alone do not necessarily reflect the quality of care provided by hospitals. Accordingly, lower MSPB-Hospital measure across performance periods (i.e., lower Medicare spending per beneficiary) in isolation should not be interpreted as better care. The MSPB-Hospital measure is most meaningful when presented in the context of other quality measures, which are part of the Hospital Value-Based Purchasing (VBP) Program. As part of the Hospital VBP Program, the MSPB-Hospital measure is aligned with current quality of care measures to facilitate profiling hospital value (payments and quality). Improvement on this measure for a hospital would be observed as a lower MSPB-Hospital measure value across performance periods.

S.12. **Detail Score Estimation** *(Detail steps to estimate measure score.)*
A hospital's MSPB-Hospital measure score is calculated as a hospital's average MSPB-Hospital amount divided by the episode-weighted median MSPB-Hospital amount across all hospitals. A hospital's MSPB-Hospital amount is defined the average spending level for the hospital's MSPB-Hospital episodes divided by the average expected episode spending level for the hospital's episodes, multiplied by the average spending over all episodes across all hospitals nationally. S.7.2. provides additional details describing the eight steps used to calculate hospitals' MSPB-Hospital measure.

**Reporting Guidelines**
This section is optional and will be available for users of the measure as guidance for implementation and reporting.

S.13.1. **Describe discriminating results approach**

Detail methods for discriminating differences (reporting with descriptive statistics--e.g., distribution, confidence intervals).
The distribution of hospitals´ MSPB Measure scores for the period of January 1, 2015 through December 31, 2015 is as follows:
Minimum: 0.47
10th Percentile: 0.88
25th Percentile: 0.94
50th Percentile: 0.99
75th Percentile: 1.03
90th Percentile: 1.08
Maximum: 2.90
This distribution of hospitals' MSPB Measure values is provided to hospitals as part of their hospital-specific reports (HSRs). Recall from S.7.2. that the denominator of the MSPB-Hospital measure is weighted by the number of episodes; as a result, the (unweighted) median MSPB-Hospital measure score is not necessarily always equal to one.

The MSPB-Hospital measure is also reported to hospitals with information about the national average measure and the state average measure, for the specific state that the hospital is a part of. Hospitals can also see the national and state average observed and expected spending per MDC and the national and state percent of spending for each claim type within the episode window. With this information, hospitals can identify the areas where the observed and expected spending are most concentrated and is most different from the national and state average.

Because CMS uses the full population of Medicare Parts A and B claims data to calculate the MSPB-Hospital measure and due to the large sample sizes, confidence intervals are of limited value. The calculated MSPB-Hospital measure represents the true measure for the time period of interest. A confidence interval is still of value in assessing the "statistical noise" in a hospital's measure score, but the reliability metrics presented in this submission also formally assess the extent of "statistical noise" and the ability to distinguish one provider's performance from another's. Further, most hospitals have a large number of episodes and thus any reported confidence intervals calculated using standard statistical methods would be fairly narrow. About 96% of hospitals have 50 or more episodes and 93% of hospitals have 100 or more MSPB-Hospital episodes.

S.13.2. **Detail attribution approach**

Detail the attribution rules used for attributing resources/costs to providers (e.g., a proportion of total measure cost or frequency of visits during the measure's measurement period) and provide rationale for this methodology.

The MSPB-Hospital episode is attributed to the hospital on the trigger inpatient claim for the index hospital admission that begins an MSPB-Hospital episode. Hospitalizations eligible to start an MSPB-Hospital episode must end in a discharge 30 days prior to the end of the period of performance to permit the collection of claim information during the post-discharge period. For example, if the period of performance is May 1, 2015 to December 31, 2015, hospitalizations must have a discharge date on or before December 1 to be eligible to be included as index admissions.

As discussed in S.9.1., however, due to the uncertainty surrounding attributing episodes to hospitals in cases where the patient was transferred between acute hospitals during the index admission, acute-to-acute transfers during the index admission are not considered index admissions for the purposes of the MSPB-Hospital measure. In other words, these cases will not generate new MSPB-Hospital episodes; neither the hospital which transfers a patient to another short-term acute hospital nor the receiving short-term acute hospital will have an index admission attributed to them.

S.13.3. **Identify and define peer group**

Identify the peer group and detail how peer group is identified and provide rationale for this methodology.

All short-term acute inpatient prospective payment system (IPPS) hospitals.

In the current MSPB-Hospital approach, only episodes triggered by short-term acute IPPS hospital claims (IPPS) are included in the measure. Only claims for beneficiaries admitted to short-term acute IPPS hospitals during the period of performance are included in the calculation of the MSPB-Hospital measure. Short-term acute IPPS hospitals are hospitals in the 50 States and D.C. other than: psychiatric hospitals, rehabilitation hospitals, and long-term care hospitals. The measure also excludes inpatient facilities whose patients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer. [1]

Although this measure was developed for public reporting and incentive payment programs for hospitals that Medicare pays under the IPPS system, one can readily expand this measure to include hospitals outside of the IPPS system, such as hospitals in Maryland and other non-IPPS hospitals.

[1] The MSPB-Hospital uses the CMS definition of a cancer hospital: http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/PPS_Exc_Cancer_Hospasp.html

S.13.4. **Sample size**

Detail the sample size requirements for reporting measure results.

The MSPB-Hospital measure will be publicly reported on Hospital Compare for hospitals with 25 or more eligible episodes. The MSPB-Hospital measure is used in the Hospital Value-Based Purchasing (VBP) Program for VBP-eligible hospitals with 25 or more eligible episodes.

S.13.5. **Define benchmarking and comparative estimates**

Detail steps to produce benchmarking and comparative estimates and provide rationale for this methodology.

The MSPB-Hospital measure can be scored against benchmarks for the purpose of inclusion in incentive payment or other performance measurement programs. In this way, value in healthcare can be recognized and incentivized. The Hospital VBP Program provides financial incentives to short-term acute hospitals based on their performance on selected quality measures. By measuring the cost of care through the MSPB-Hospital measure, CMS aims to recognize hospitals that can provide high quality care at a lower cost to Medicare. Combined with the other quality measures that comprise the Total Performance Score (TPS) under the Hospital VBP Program, the MSPB-Hospital measure allows CMS to assess the value of care and incentivize both achievement and improvement in efficiency.

Under the Hospital VBP Program, hospital performance on the MSPB-Hospital measure will be determined using the higher of its

achievement or improvement score, as described in the FY 2012 IPPS Final Rule at 76 FR 51654-56.  The MSPB-Hospital measure score will then be included in the hospital's Total Performance Score (TPS) within the Efficiency domain.
For information on how the MSPB-Hospital measure score will be incorporated into the Hospital VBP Program, please refer to the FY 2012 IPPS/LTCH PPS final rule: http://www.gpo.gov/fdsys/pkg/FR-2011-08-18/pdf/2011-19719.pdf.

**Validity – See attached Measure Testing Submission Form**

**SA.1. Attach measure testing form**
2016_11_04_mspb_hospital_nqf_testing_form.docx

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

**Measure Number** (*if previously endorsed*)**:** 2158
**Measure Title**: Medicare Spending Per Beneficiary (MSPB) - Hospital
**Date of Submission**: 11/4/2016
**Type of Measure:**

| | |
|---|---|
| ☐ Outcome (including PRO-PM) | ☐ Composite – STOP – use composite testing form |
| ☐ Intermediate Clinical Outcome | ☒ Cost/resource |
| ☐ Process | ☐ Efficiency |
| ☐ Structure | |

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. If there is more than one set of data specifications or more than one level of analysis, contact NQF staff about how to present all the testing information in one form.
- For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.
- For outcome and resource use measures, section 2b4 also must be completed.
- If specified for multiple data sources/sets of specificaitons (e.g., claims and EHRs), section 2b6 also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for supplemental materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (incuding questions/instructions; minimum font size 11 pt; do not change margins). Contact NQF staff if more pages are needed.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For PRO-PMs and composite performance measures, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For PRO-PMs and composite performance measures, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the

exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful [16] differences in performance;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For eMeasures, composites, and PRO-PMs (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

**NOTE**

Acumen is submitting the Medicare Spending Per Beneficiary (MSPB) - Hospital measure for National Quality Forum (NQF) endorsement. This document presents Acumen's responses to the NQF Testing Attachment questions for the MSPB-Hospital Measure. A supplementary Excel file titled "*2016_11_02_mspb_hospital_testing_appendix_tables.xlsx*" provides detailed results for many of the analyses summarized in this testing attachment form.

Please note that text from our previous submission is included in grey italics font type.

## DATA/SAMPLE

### 1. Data/Sample Used for All Testing of Measure

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

### 1.1. Type of Data

**What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

| **Measure Specified to Use Data From:**<br>**(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
|---|---|
| ☐ abstracted from paper record | ☐ abstracted from paper record |
| ☒ administrative claims | ☒ administrative claims |
| ☐ clinical database/registry | ☐ clinical database/registry |
| ☐ abstracted from electronic health record | ☐ abstracted from electronic health record |
| ☐ eMeasure (HQMF) implemented in EHRs | ☐ eMeasure (HQMF) implemented in EHRs |
| ☒ other: Long-term Minimum Data Set and Enrollment Database | ☒ other: Long-term Minimum Data Set, Enrollment Database, and American Community Survey |

### 1.2. Dataset

**If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Medicare Parts A and B claims data from the Common Working File (CWF), Long-term Minimum Data Set (MDS) data, Enrollment Database (EDB) data, and the United States Census Bureau's American Community Survey.

***Previous response:***

*Medicare Parts A and B claims data from the Common Working File (CWF).*

### 1.3. Date Range

**What are the dates of the data used in testing**?

Inpatient admissions with a discharge date between January 1, 2015 and December 1, 2015.

For the test-retest analysis, data also included inpatient admissions with a discharge date between January 1, 2014 and December 1, 2014.

***Previous response:***

*May 15, 2010 – February 14, 2011*

### 1.4. Levels of Analyses Tested

**What levels of analysis were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

| **Measure Specified to Measure Performance of:** <br> **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
|---|---|
| ☐ individual clinician | ☐ individual clinician |
| ☐ group/practice | ☐ group/practice |
| ☒ hospital/facility/agency | ☒ hospital/facility/agency |
| ☐ health plan | ☐ health plan |
| ☐ other: Click here to describe | ☐ other: Click here to describe |

### 1.5. Measured Entities

**How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

3,298 Inpatient Prospective Payment System (IPPS) hospitals with discharges between 1/1/2015 and 12/1/2015 received an MSPB-Hospital measure value. Only claims for beneficiaries admitted to subsection (d) hospitals during the period of performance are included in the calculation of the MSPB-Hospital measure. Subsection (d) hospitals are hospitals in the 50 States and D.C. other than: psychiatric hospitals, rehabilitation hospitals, hospitals whose inpatients are predominantly under 18 years old, hospitals whose average inpatient length of stay exceeds 25 days, and hospitals involved extensively in treatment for or research on cancer.

***Previous response:***

*3,396 IPPS hospitals received an MSPB Measure value (5/15/2010-2/14/2011 period of performance)*

### 1.6. Patient Population

**How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

4,261,069 beneficiaries (from 5,531,258 episodes) were included in the testing and analysis. These beneficiaries are enrolled in Medicare fee-for-service and were discharged from short-term acute hospitals between 1/1/2015 and 12/1/2015. Specifically, Medicare Part A and Medicare Part B claims from beneficiaries with an index admission within a subsection (d) hospital are included in the MSPB-Hospital episode if the beneficiary has been enrolled in Medicare Part A and Part B for the period 90 days prior to the start of an episode (i.e., 93 days prior to the date of the index admission) until 30 days after discharge.

To determine whether the MSPB-Hospital measure inclusion criteria distort patient characteristics on index admissions, we produced and analyzed distributions of patient characteristics (age, race, and sex) for two groups of patients: one group in which the beneficiaries had an eligible admission, and the other group in which patients both had an eligible admission and met the specified inclusion criteria as specified above. Appendix Tables 1-1, 1-2, and 1-3 detail these distributions and show that the MSPB-Hospital measure inclusion criteria do not significantly change the percentage of beneficiaries of any particular demographic. The typical difference between groups for a given characteristic is usually within 1 percentage point. To illustrate, the percent of beneficiaries aged 70 to 75 in the group that applies the inclusion criteria is 17%, compared to 16% when not implementing the inclusion criteria. The breakdown of race (i.e., Black and Non-Black) with and without the inclusion criteria is nearly identical. The breakdown of male and female beneficiaries with and without the inclusion criteria is also very similar, as the composition is 56% female in the group implementing the inclusion criteria compared to 55% when not applying the inclusion criteria.

***Previous response:***

*3,566,422 beneficiaries. These beneficiaries are enrolled Medicare fee-for-service and were discharged from short-term acute hospitals between (5/15/2010 and 2/14/2011)*

### 1.7. Differences in Data Used in Different Aspects of Testing

**If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

N/A. The data samples used for the different aspects of testing below are identical. The test-retest analysis looked at data from one year prior as well, as noted in Section 1.3.

***Previous response:***

*The data samples used for the different aspects of testing below are identical.*

### 1.8 SES Variables

What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

The socioeconomic (SES) factor we analyzed is family income-to-poverty ratio. We obtained community-level poverty data from the 2014 American Community Survey, accessed through the United States Census Bureau's American FactFinder website, to determine the number of families in a given ZIP code whose income-to-poverty ratio (the ratio of family income to the federal poverty threshold) falls into certain categories. The dataset "Ratio of Income to Poverty Level of Families in the Past 12 Months" contains variables that represent ranges of income-to-poverty ratios. The values for these variables are the number of families in a given ZIP code whose income-to-poverty ratio falls into that variable's income-to-poverty ratio range. For example, if the value for the ".50 to .74" variable is 10,000 for a particular ZIP code, that means that 10,000 families in that ZIP code have incomes that are between 50% and 74% of the federal poverty threshold.

Enrollment Database (EDB) data provided the ZIP codes for beneficiaries included in the sample. We then linked these beneficiary ZIP codes to the ACS ZIP code-level data on family income-to-poverty ratio, which allowed us to analyze poverty data in beneficiaries' ZIP codes. We used family income-to-poverty ratio instead of individual income-to-poverty ratio to better reflect actual financial assets available to beneficiaries, as individual family members may pool financial resources to provide care for older relatives.

## 2A2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

### 2a2.1. Level of Reliability Testing

**What level of reliability testing was conducted**? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

### 2a2.2. Method

**For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

*Data Element Reliability*:

To construct the MSPB-Hospital measure, Acumen uses CMS claims data. CMS has in place several hospital auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in our measures, including diagnosis and procedure codes and other elements that are consequential to payment. Specifically, CMS works with Program Safeguard Contractors (PSCs)/Zone Program Integrity Contractors (ZPICs) to ensure program integrity; the agency also uses Comprehensive Error Rate Testing (CERT) Contractors to ensure that Medicare payments are correct. Between 2005 and 2015, CERT estimates that proper payment, which is payments that met Medicare coverage, coding, and billing rules, ranged from 87.3 to 96.4 percent of total payments each year.[1] CMS continues to perform successful corrective actions and give providers additional education to ensure accurate billing. To ensure claims completeness and inclusion of any corrections, the measure is calculated using data with a 3 month claims run-out from the end of the performance period.

*Measure Reliability*:

Measure reliability is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. To estimate measure reliability, we utilize two approaches: (1) Test/Retest and (2) Reliability Score.

[1] Comprehensive Error Rate Testing (CERT) Program. "Appendices Medicare Fee-for-Service 2015 Improper Payments Report". Table A6. https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Medicare-FFS-Compliance-Programs/CERT/CERT-Reports-Items/Downloads/AppendicesMedicareFee-for-Service2015ImproperPaymentsReport.pdf

Our first approach to assess reliability is to consider the extent to which assessments of a hospital using unique sets of episodes produce similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured using two sets of episodes. We examine the correlation and quintile rank stability between a hospital's MSPB-Hospital scores calculated from both samples. By comparing the correlation of a hospital's MSPB measure calculated using the two mutually exclusive samples, one can identify the relationship of a hospital's score across samples. For this analysis, Acumen performed two separate test/retest investigations: comparing two random subsets of episodes from 2015, and comparing the set of 2015 episodes to the set of 2014 episodes. Both investigations sought to identify the reliability of a hospital's score across samples.

Our second approach calculates reliability scores as: $R_j = V_b/(V_b + (V_{w_j}/n_j))$ where $R_j$ is the reliability for hospital *j*, $V_b$ is the between hospital variance, $V_{w_j}$ is the within hospital variance for hospital *j*, and $n_j$ is the number of MSPB episodes for hospital *j*. This analysis seeks to determine the extent to which variation in the measure is due to true, underlying hospital performance rather than random variation (i.e. statistical noise) within hospitals due to the sample of cases observed.

***Previous response:***

*Data Element Reliability: Due to CMS's extensive auditing program, we believe that patient demographics, diagnostic information, and payment information are very reliable. As described in F.4., CMS uses various auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and for overpayment recoupment. CMS also routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in our measures.*

*Measure Reliability: The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is naturally the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. To estimate measure reliability, we utilize four approaches: (1) Test/Retest, (2) Seasonality, (3) Reliability Score, and (4) Bootstrapping.*

*Our first approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produces similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured once using a random subset of patients, then measured again using a second subset (over the same time period) that excludes the MSPB episodes chosen for the first sample. We examine the correlation, and quintile rank stability between a hospital's MSPB scores calculated from both samples.*

*Second, because the MSPB Measure values reported on Hospital Compare in April 2012 use Medicare claims data from May through February, Acumen conducted a seasonality analysis to examine how MS-DRGs change within a year. Providers that efficiently treat specific DRGs may receive higher MSPB Measure values during a season where the DRG occurs frequently and lower MSPB Measure values during a season where the DRG occurs less frequently. For this specific analysis, we split inpatient claims data with through date in 2010 into two categories: claims with through dates from January through April and claims with through dates from May through December.*

*Our third approach calculates reliability scores as:* $R_j = V_b/(V_b + (V_{w_j}/n_j)$ *where* $R_j$ *is the reliability for Hospital j,* $V_b$ *is the between hospital variance,* $V_{w_j}$ *is the within hospital variance for hospital j, and* $n_j$ *is the number of MSPB episodes for hospital j.*

*Fourth, Acumen measured how reliability varies based on the number of MSPB episodes a hospital is assigned. This fourth analysis is divided into two parts. The first evaluates how the number of MSPB episodes a hospital receives affects its 95 percent confidence interval. This analysis also informs how CMS should set the minimum number of episode required for public reporting purposes. When increasing the threshold for the minimum number of cases (or hereafter referred to as 'episode'), one decreases the likelihood an outlier episode[2] materially affects a hospital's MSPB score, but also decreases the number of hospitals able to publicly report their MSPB Measure.*

*Whereas determining the number of hospitals that would be dropped when the minimum episode threshold increases is straight-forward, our second approach for measuring the effect of the minimum episode threshold on the MSPB confidence interval requires additional explanation. Typically, confidence intervals are constructed for commonly used quantities, such as the sample mean in which the distribution of the sample quantity is known, and can be used in the interval calculation. However, the MSPB score is a ratio of weighted means and does not have an easily identifiable statistic that corresponds to dispersion. Further, the MSPB score is not normally distributed, and typical measures of the dispersion of a distribution—such as the standard deviation—will not fully characterize the variation in the MSPB distribution.*

*In this analysis, Acumen instead uses a non-parametric bootstrap methodology to measure how the confidence interval of the MSPB score changes when the minimum episode threshold increases. This analysis measures the MSPB score for an 'average' hospital, where the 'average' hospital case is considered to be one whose MSPB episode distribution mimics that of the entire population of MSPB episodes. The bootstrap simulates the process of randomly drawing MSPB episodes from the population, and thus approximates the actual shape of the MSPB score distribution from which confidence intervals are determined. By repeatedly calculating an MSPB score for this simulated hospital under differing assumptions on the number of episodes observed, one can create a confidence interval for the MSPB score of this 'average' hospital.*

*To implement the bootstrap procedure, this analysis examines cases where the 'average' hospital has X episodes, where X = 1, 2, 3, 5, 10, 25, and 100. The five step methodology used to implement this analysis is as follows: (1) Draw 10,000 random samples (with replacement) each with X number of episodes from the original dataset containing MSPB episodes; (2) Calculate MSPB Amount for each sample; (3) Calculate MSPB Measure—normalization of the MSPB Amount—as the MSPB Amount for the hospital divided by the median MSPB Amount across all hospitals; (4) Calculate the 95 percent confidence interval using the 2.5th and 97.5th percentiles*

[2] Statistical outlier episodes are excluded from the MSPB calculation to mitigate the effect of high-cost and low-cost outliers on each hospital's MSPB Measure. The MSPB Measure methodology uses "residuals" to define outlier episodes, where a residual equals the standardized episode spending minus the expected episode spending. High-cost outliers are defined as episodes whose residual falls above the 99th percentile of the residual cost distribution within any MS-DRG admission category; similarly, low-cost outliers are defined as episodes whose residual falls below the 1st percentile of the residual cost distribution within any MS-DRG category. For additional details on the definition of statistical outliers for the MSPB Measure, see the response to Question 2a1.20 of this measure submission form.

*of the MSPB Measure distribution;[3] and (5) Divide the width of this confidence interval by the width of the confidence interval for X = 100 episodes.*

### 2a2.3. Results

**For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

1. *Test/Re-Test*: For the 2014 and 2015 sample (i.e., comparing 2015 data to 2014 data), over 75 percent of hospitals in the lowest-spending quintile in one year are in the lowest-spending quintile in the other; similarly, over 74 percent of hospitals in the highest-spending quintile in one year are in the highest-spending quintile in the other. Moreover, over 91 percent of hospitals in the highest-spending quintile in one year are in one of the top two highest spending quintiles in the other year. Quintiles results are listed in Appendix Table 2a2-1. The Spearman rank correlation for a hospital across the two years is 0.85, and the Pearson correlation coefficient is 0.81. As a point of comparison, in a standard moving-average time series process with one lag (i.e., an MA(1) process), the maximum possible Pearson correlation is 0.50.[4] Therefore, the value of 0.81 is remarkably high in relation to a relevant statistical benchmark.

   For the 2015 sample (i.e., comparing two random subsets of episodes from 2015), over 72 percent of hospitals in the lowest-spending quintile in one sample are in the lowest-spending quintile in the next; similarly, over 71 percent of hospitals in the highest-spending quintile in one sample are in the highest-spending quintile in the next. Moreover, over 90 percent of hospitals in the highest-spending quintile in one sample are in one of the top two highest spending quintiles in the next. The Spearman rank correlation for a hospital across samples is 0.82, and the Pearson correlation coefficient is 0.70. In a simple econometric model where two outcomes share a mean and each have two additive error terms (one in common, and one distinct), the Pearson correlation is 0.50. [5] The value of 0.70 is high relative to this statistical benchmark in which the expected value of the two outcomes are completely identical.

2. *Reliability Score*: Using a minimum episode threshold of 25 MSPB-Hospital episodes, over 99 percent of hospitals have a reliability score greater than 0.4 and 67.9 percent of hospitals have a reliability score greater than 0.9. Additionally, the average reliability score for hospitals with at

[3] If a hospital has a true MSPB Measure value of 1.0, a 95% confidence interval indicates that 95% of the time the hospital's MSPB Measure value will fall between the 2.5th and 97.5th percentiles if the hospital gets *X* number of episodes from the original dataset containing MSPB episodes.

[4] Goldberger, 1991, *A Course in Econometrics*, and Greene, 2002, *Econometric Analysis*. An MA(1) model of a dependent variable such as the MSPB score takes the form $y_t = \mu + u_t + \theta u_{t-1}$, where *t* indicates the time period, $\mu$ is a constant over time, and $u_t$ and $u_{t-1}$ are mean zero, independent error terms.

[5] This example parallels the MA(1) time series example in footnote 2; see the references there for details. The econometric model of two outcomes in time period $t$ , $y_{t1}$ and $y_{t2}$, is given by $y_{t1} = \mu + \epsilon_t + u_{t1}$ and $y_{t2} = \mu + \epsilon_t + u_{t2}$, where $\mu$ is the shared mean, and $\epsilon_t$ , $u_{t1}$ and $u_{t2}$ are independent, mean zero error terms with common variance.

least 25 episodes is 0.897. Previous work supported that 0.4 is the lower limit of "moderate" reliability;[6] the MSPB-Hospital measure exceeds this threshold for over 99 percent of hospitals.

***Previous response:***

*1. Test/Re-Test: Over 70 percent of hospitals in the lowest-spending quintile in one sample are in the lowest-spending quintile in the next; similarly, over 70 percent of hospitals in the highest-spending quintile in one sample are in the highest-spending quintile in the next. The Spearman rank correlation for a hospital across samples is 0.835.*

*2. Seasonality Analysis: Between the January 2010 – April 2010 period and the May 2010 – December 2010 period, the average absolute change in the relative frequency of an MS-DRG index admission was 8.9%. Certain lung-related admissions (e.g., pneumonia, COPD, asthma) appear more frequently in the winter.*

*3. Reliability Score: The MSPB Measure's overall reliability is 0.951. Over 98 percent of hospitals have a reliability score greater than 0.4; 62 percent of hospitals have a reliability score greater than 0.9. Previous work proposed that 0.4 is the lower limit of "moderate" reliability;[7] the MSPB measure exceeds this threshold.*

*4. Minimum Number of Cases Required for the MSPB Measure: As the minimum episode threshold increases, there is a trade-off between the size of the confidence interval for the 'average' hospital and the number of hospitals receiving an MSPB score. Table 1 in the appendix shows that as the minimum episode threshold, X, increases, the confidence interval becomes narrower and more reliable. Specifically, the 95% confidence interval decreases by almost a third as cutoff number is moved from X = 5 to X = 50. However, as the minimum episode threshold increases from X = 5 to X = 50, the number of hospitals that could publicly report this measure included decreases; in fact, at the cutoff X = 50 episodes, the share of hospitals included decreases to 95.9%.*

### 2a2.4. Interpretation

**What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

1. *Test/Retest*: Sample selection does not have a material effect on a hospital's MSPB-Hospital measure for different data samples drawn from the same period, or for data samples drawn from different periods. . In other words, hospitals have similar MSPB-Hospital measure quintile ranks regardless of which MSPB-Hospital episodes are used to calculate the MSPB-Hospital measure scores. This indicates that the MSPB-Hospital measure score is a reliable measure of a hospital's risk-adjusted Medicare spending compared to other hospitals.

[6] Mathematica, Inc. "Memorandum: Reporting Period and Reliability of AHRQ, CMS 30-Day and HAC Quality Measures – Revised." http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/HVBP_Measure_Reliability-.pdf

[7] Mathematica, Inc. "Memorandum: Reporting Period and Reliability of AHRQ, CMS 30-Day and HAC Quality Measures – Revised." http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/HVBP_Measure_Reliability-.pdf

2. *Reliability Score*: Overall reliability of the MSPB-Hospital measure is extremely high due to the large number of MSPB-Hospital episodes attributed to most hospitals. Reporting the MSPB-Hospital measure for hospitals that have at least 25 attributed episodes provides a balance between reliability and measure inclusiveness.

***Previous response:***

*1. Quintile Rank Stability Across Groups: Sample selection does not have a material effect on a hospital's MSPB score for different data samples drawn from the same period.*

*2. Seasonality Analysis: The seasonality analysis indicates that the incidence of different types of hospitalizations (i.e., MS-DRGs) varies across the year, but this variability for the most part is concentrated in DRGs lung-related diseases.*

*3. Reliability Score: Overall reliability of the MSPB score is extremely high due to the large number of MSPB episodes attributed to most hospitals. Reporting the MSPB Measure for hospitals that have at least 25 attributed episodes provides a balance between reliability and measure inclusiveness.*

*4. Minimum Number of Cases Required for the MSPB Measure: Based on the empirical results presented in 2a2.3., reporting the MSPB Measure as part of the Hospital VBP program for hospitals that have at least 25 attributed episodes provides a balance between the size of the confidence interval and the number of hospitals receiving and MSPB Measure score.*

## 2B2. VALIDITY TESTING

### 2b2.1. Level of Validity Testing

**What level of validity testing was conducted**? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

### 2b2.2. Method

**For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Acumen utilized three tests to evaluate the validity of the MSPB-Hospital measure: (1) correlation with another measure of Medicare spending, specifically CMS' measure of risk-adjusted, standardized total Medicare spending at the Hospital Referral Regions (HRR) level, (2) correlation with service utilization rates, and (3) cost variation by time period. The first two correlations seek to confirm the validity of the MSPB-Hospital measure by comparing it with other measures of resource use, while the third test seeks to confirm the measure's validity by determining if cost variation by time period is consistent with expectations.

The first test examined the correlation between the MSPB-Hospital measure and the measure of risk-adjusted, aggregated annual per-capita spending for all Medicare beneficiaries produced by CMS at the HRR level.[8] This measure included all Medicare beneficiaries that had no months of Medicare Advantage enrollment and had both Part A and Part B for the portion of the year that they were covered by Medicare. Data on this measure of Medicare spending were available for 2007 – 2014, and Acumen performed correlation analyses for each of those years. For each HRR, Acumen found the mean MSPB-Hospital measure and correlated with the risk-adjusted, standardized, per capita HRR-level measure of total Medicare spending. This analysis sought to confirm the accuracy of the MSPB-Hospital measure by comparing its findings to a measure of Medicare spending.

The second test examined the correlation between the MSPB-Hospital measure and a measure of service utilization constructed by Acumen. To construct the service utilization measure, Acumen constructed hospital-level averages of services billed during the MSPB-Hospital episode across various categories (professional Evaluation & Management (E&M), post-acute, etc.). Acumen subsequently correlated these averages with the MSPB-Hospital measure. This analysis sought to confirm the expectation that the MSPB-Hospital measure correlates with service utilization rates.

The third test examined cost variation by time period. To do so, we broke down the total variance in risk-adjusted cost by time period, namely the period 3 days prior to and during the index admission and the period post-discharge. Because the risk adjustment model controls for MS-DRG, and because the MS-DRG of the index admission is the primary driver of costs from 3 days prior and during the index admission, the expected result of this analysis is that risk-adjusted episode cost should be strongly driven by post-discharge cost.

***Previous response:***

*The first validity test examines the correlation between hospitals' MSPB scores and the percent of beneficiaries with multiple episodes. This analysis examines whether high-cost hospitals may have below average (i.e., efficient) MSPB Measure values if the MSPB episode definition separates a single episode of care into two or more MSPB episodes. Division of a single episode of care into multiple MSPB episodes occurs when a hospital admission takes place more than 30 days after the initial discharge.*

*The second test of the validity of the MSPB Measure compares the MSPB Measure against other related outcome measures. Specifically, we will examine whether hospitals with low MSPB scores (i.e., efficient hospitals) are also less likely to have various types of hospital readmissions.*

**2b2.3. Results**

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

1. *Correlation with Another Measure of Medicare Spending:* For each year for which the risk-adjusted, standardized, per capita HRR-level measure data were available (2007 to 2014), the MSPB-Hospital measure had a positive correlation of at least 0.5 with the corresponding HRR-level measure. From 2007 to 2014, the lowest Spearman rank correlation for a given year was

[8] Centers for Medicare & Medicaid Services. "Medicare Geographic Variation Public Use File." http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Geographic-Variation/GV_PUF.html

0.53 and the lowest Pearson correlation coefficient was 0.51; during the same period, the highest Spearman rank correlation was 0.63 and the highest Pearson correlation coefficient was 0.61.

2. *Correlation with Service Utilization Rates:* The MSPB-Hospital measure had a Pearson correlation of 0.42 with professional E&M services per episode and a Pearson correlation of 0.52 with post-acute skilled nursing and inpatient services per episode.

3. *Cost Variation by Time Period:* For the MSPB-Hospital measure, costs during the post-discharge period account for over 84 percent of total MSPB-Hospital episode cost variance, while costs from the period 3 days prior to and during the index admission account for just over 11 percent of total episode cost variance. These results are also shown in Appendix Table 2b2-1.

***Previous response:***

*1. Beneficiaries with Multiple Episodes: The analysis indicated a positive correlation between MSPB Measure values and the percent of beneficiaries with multiple episodes. The hospital-level correlation between the MSPB Measure and the percent of beneficiaries with multiple episodes was 0.13; when accounting for variation in the MS-DRG of the index admission when measuring readmission rates, the correlation between readmissions and the MSPB Measure increases slightly to 0.16.*

*2. Correlation with Other Outcome Measures: The MSPB Measure exhibits a positive correlation with a number of hospital readmission measures. The correlation between the MSPB Measure and Heart Attack, Heart Failure, and Pneumonia Readmission Rates are of 0.08, 0.07, and 0.06, respectively.*

**2b2.4. Interpretations**

**What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The interpretation of correlation results can depend on the specific analysis. In a simple econometric model where two outcomes share a common mean with additive and identically distributed errors, the Pearson correlation is 0.5 (see previous footnotes in the reliability testing Section 2a2.3).[9]

1. *Correlation with Another Measure of Medicare Spending:* The positive correlation between the MSPB-Hospital measure and the risk-adjusted, standardized, per capita HRR-level measure of Medicare spending indicates that the MSPB-Hospital measure's identification of hospitals with high- or low risk-adjusted spending is consistent with a measure of Medicare spending.

2. *Correlation with Service Utilization Rates:* The positive correlation between the MSPB-Hospital measure and service utilization rates, specifically for E&M services and post-acute nursing and inpatient services, indicates that the MSPB-Hospital measure accurately captures higher resource use.

[9] Goldberger, 1991, *A Course in Econometrics*, and Greene, 2002, *Econometric Analysis*.

3. *Cost Variation by Time Period:* Variance in costs during the post-discharge period makes up a larger portion of total variance than variance in costs during the period 3 days prior to and during the index admission does. This finding is consistent with expectations. The risk adjustment model predicts a certain level of post-discharge spending based upon the beneficiary's prior health history and MS-DRG. This analysis shows that of the cost variance left over after this risk adjustment, most of it is driven by post-discharge spending. Variance in provider scores based on post-discharge spending emphasizes the importance of care transitions and care coordination in improving patient care.

***Previous response:***

*1. Beneficiaries with Multiple Episodes: Hospitals are not likely to be postponing necessary re-admissions—and thus creating a new episode—to improve their MSPB Measure values. High-cost hospitals are not more likely to treat beneficiaries with multiple hospitalization episodes.*

*2. Correlation with Other Outcome Measures: The positive correlation between the MSPB Measure and Heart Attack, Heart Failure, and Pneumonia Readmission Rates indicate that hospitals that are more expensive generally have higher readmission rates. The correlation, however, is weak for all three readmission rates. A weak correlation can be explained by the fact that the MSPB Measure assesses the cost to Medicare of all services performed by hospitals and other healthcare providers during an MSPB episode. As a result, a hospital's MSPB Measure value is driven by both acute and post-acute spending.*

## 2B3. EXCLUSIONS ANALYSIS

### 2b3.1. Method

**Describe the method of testing exclusions and what it tests** (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Acumen evaluated the validity of the measure exclusion criteria by producing impact analyses, which show the effect of recalculating the MSPB-Hospital measure while independently reversing each of the following exclusion criteria: (1) acute-to-acute transfer episodes;[10] (2) death episodes;[11] and (3) outlier episodes.[12] For (1), our analysis evaluated the impact of including transfer episodes on MSPB-Hospital measure scores. For (2), we re-calculated the MSPB-Hospital measure using beneficiaries who die during the episode. Specifically, we examined the percent of beneficiaries who die during the MSPB-Hospital episode and the effect

[10] Transfers, defined based on the claim discharge code, are not considered eligible as index admissions. In other words, these cases will not generate new MSPB-Hospital episodes; neither the hospital which transfers a patient to another short-term acute hospital nor the receiving short-term acute hospital will have an index admission attributed to them.

[11] Recall from S.9.1. that any episode where at any time during the episode the beneficiary dies is excluded from the MSPB-Hospital calculation.

[12] Recall from S.9.1. that MSPB-Hospital episodes whose relative scores fall above the 99th percentile or below the 1st percentile of the distribution of residuals are excluded from the MSPB-Hospital calculation.

that including death episodes had on hospital scores. For (3), we examined the effect of including outliers when calculating MSPB-Hospital measure scores instead of excluding outliers based on the distribution of residuals. Specifically, we examined the impact of top-coding episodes with risk-adjusted costs that are above the 99th percentile, where those episodes are assigned the cost of the episode at the 99th percentile. We also examined the impact of bottom-coding episodes with risk-adjusted costs that are below the 1st percentile, where those episodes are assigned the cost of the episode at the 1st percentile.

The measure also implements an exclusion criteria specific to inpatient admissions that are allowed to trigger a new MSPB-Hospital measure. Specifically, we do not allow inpatient admissions that occur within 30 days post-discharge of another inpatient admission to start a new MSPB-Hospital episode; we refer to this criteria as excluding overlapping episodes. For this exclusion (4), we analyzed the effect of including overlapping episodes when constructing the MSPB-Hospital episodes. To illustrate what this exclusion is, take an inpatient admission that triggers Episode A and see if the beneficiary has another inpatient admission within the 30-day post-discharge window of Episode A. If the beneficiary has a second qualifying admission within the 30-day post-discharge window of Episode A, do not allow the second admission to trigger Episode B. We evaluated the impact of this exclusion on MSPB-Hospital measures by re-calculating MSPB-Hospital with the previously-excluded episodes added back in, which was then compared to MSPB-Hospital measures calculated under the overlapping episodes exclusion.

***Previous response:***

*Acumen evaluated the validity of the inclusion/exclusion criteria by producing impact analyses which show the effect of recalculating the MSPB Measure while independently reversing each of the following inclusion/exclusion criteria: (1) beneficiaries in Medicare Advantage; (2) beneficiaries in Medicare Part A only; (3) acute-to-acute transfers;[13] (4) death episodes;[14] and (5) outlier episodes.[15] With respect to (3), Acumen's analysis evaluates assigning transfers to the transferring hospital and to the receiving hospital. The first three restrictions occur because of incomplete data or problems attributing episodes to individual hospitals. For (4), we re-calculate the MSPB Measure using beneficiaries who die during the episode. Specifically, Acumen examined the percent of beneficiaries who die during the MSPB episode and after the MSPB episode and whether or not to calculate separate MSPB Measures for beneficiaries who died during the episode versus beneficiaries who did not die. For (5), we examine top-coding/bottom-coding distribution outliers in place of completely excluding them.*

*Acumen also conducted a number of analyses on potential exclusion criteria. These unimplemented exclusions include: (6) beneficiaries discharged against medical advice (AMA) and (7) dual-eligibles. Acumen's analysis evaluates not counting admissions in which the*

[13] Recall from S.9.1. that transfers, defined based on the claim discharge code, are not considered eligible as index admissions. In other words, these cases will not generate new MSPB episodes; neither the hospital which transfers a patient to another short-term acute hospital, nor the receiving short-term acute hospital will have an index admission attributed to them. The rationale for exclusion of these acute-to-acute transfer cases is that CMS wished to perform further analysis of hospital impacts and explore potential unintended consequences of attribution of the MSPB episode to either the transferring or the receiving hospital.

[14] Recall from S.9.1. that any episode where at any time during the episode the beneficiary becomes deceased is excluded from the MSPB calculation.

[15] Recall from S.9.1. that MSPB episodes whose relative scores fall above the 99th percentile or below the 1st percentile of the distribution of residuals (see 2a1.20 for a description of MSPB residuals) within each index admission MS-DRG are excluded from the MSPB calculation.

*beneficiary was discharged AMA as an index admission. Although excluding patients discharged against medical advice would avoid attributing the costs of non-compliant beneficiaries to a hospital's MSPB Measure value, hospitals would be incentivized to encourage high-cost beneficiaries to leave against medical advice to avoid having their episode included in the hospital's MSPB Measure. We also evaluate (i) including a dual-eligible indicator in the MSPB risk-adjustment and (ii) examining MSPB scores separately for duals/non-duals.*

**2b3.2. Results**

**What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

1. *Transfer Episodes:* Episodes that include an acute-to-acute transfer account for 1.6% of total episodes. Episodes containing an acute-to-acute transfer have an average observed cost of $33,363 compared to an average expected cost of $21,068, resulting in an observed-to-expected cost ratio of 1.58. Episodes not containing an acute-to-acute transfer, on the other hand, have an average observed cost of $20,570 compared to an average expected cost of $20,774, resulting in a observed-to-expected cost ratio of 0.99 (Appendix Table 2b3-1). Rural hospitals tend to have a higher rate of transfers than urban hospitals (4.1% and 1.3%, respectively), so including transfer episodes that have higher observed-to-expected cost ratio in the MSPB-Hospital measure calculation would probably disproportionately worsen rural hospitals' scores. When including transfer episodes in the calculation of the MSPB-Hospital measure, 81% of hospitals' MSPB-Hospital measure scores change by less than ±0.03, and less than 2% of hospitals' MSPB-Hospital measure scores change by more than ±0.10 (see Appendix Table 2b3-2 for full results). The correlation between MSPB-Hospital measure scores when excluding transfer episodes versus when including transfer episodes is 0.95.

2. *Death Episodes:* In approximately 8% of MSPB-Hospital episodes, the beneficiary dies before the end of the 30-day post-discharge period. Episodes in which the beneficiary dies during the episode window (denoted as "death episodes") appear more efficient than non-death episodes, as shown in Appendix Table 2b3-3. The average observed cost of death episodes is $21,041 compared to the expected cost of $24,980, resulting in an observed-to-expected cost ratio of 0.84. Comparatively, non-death episodes have an observed-to-expected cost ratio of 1.02 ($20,512 over $20,156). If death is included in measure calculation, 96% of hospitals' MSPB-Hospital measure scores change by less than ±0.03, and very few hospitals (less than 0.2%) see changes in MSPB-Hospital measure scores greater than ±0.10 (see Appendix Table 2b3-4). The correlation between MSPB-Hospital measure scores when excluding death episodes versus when allowing for inclusion of death episodes in measure calculation is 0.99.

3. *Outlier Episodes:* When including outlier episodes in measure calculation, about 2% of hospitals see an absolute change in their MSPB-Hospital measure score of greater than ±0.10, and 6% of hospitals' MSPB-Hospital measure scores change by greater than ±0.05.

Appendix Table 2b3-5 further details the impact of including outliers on MSPB-Hospital measure scores. The correlation between MSPB-Hospital measure scores when excluding outliers versus when including outliers is 0.93.

4. *Overlapping Episodes:* Approximately 12% of episodes had their trigger inpatient admission within 30 days of the discharge date of the trigger inpatient admission of another episode (Appendix Table 2b3-6). If episodes with a trigger inpatient admission during the 30-day post-discharge period of another episode are included in MSPB-Hospital measure calculation, 97% of hospitals' MSPB-Hospital measure scores change by less than ±0.03, with a small proportion of hospitals (0.4%) experiencing changes in MSPB-Hospital measure scores greater than ±0.10 (see Appendix Table 2b3-7 for detailed results). The correlation of MSPB-Hospital measure scores before and after removing the overlapping episodes exclusion is 0.99.

***Previous response:***

*Medicare Advantage or Part A Only: 25% of Medicare beneficiaries are enrolled in Medicare Advantage; about 10 percent of Medicare FFS beneficiaries are enrolled in Part A only.*

*Transfers: Episodes that include an acute-to-acute transfer account for 5% of total episodes. Episodes containing an acute-to-acute transfer have an average risk-adjusted spending of $25,151 per episode, while the average episode not containing an acute-to-acute transfer has an average risk-adjusted spending of $19,489 per episode. Because transfer episodes cost 29% more than non-transfer episodes on average, excluding transfer episodes eliminates a significant portion of MSPB episodes and Medicare payments. Small rural hospitals are the most likely facilities to transfer to large, urban hospitals (see Tables 2 and 3 in the appendix). Assigning transfer episodes to the transferring hospital has a larger effect on the MSPB Measure than assigning transfer episodes to the receiving hospital. When transfer episodes are assigned to the receiving hospital, 90% of hospitals experience a change in their MSPB Measure values of less than 3 percent, but only 80% of hospitals experience a change in their MSPB Measure values of less than 3 percent when transfer episodes are assigned to the transferring hospital (see Tables 4 and 5 in the appendix)*

*Death Episodes: In approximately 8.0% of MSPB episodes, the beneficiary dies before the end of the 30-day post-acute period. Death episodes are much more expensive than non-death episodes. Whereas death episodes cost $26,883 on average, non-death episodes cost $19,141, a 40% difference in average episode cost. Since death episodes are typically expensive, including death episodes in the MSPB Measure would increase the skewness of the episode cost distribution. Including death episodes (after outlier episodes have been excluded) increases the ratio of the 99th percentile cost to the median cost by 3 percent. If death is included as a variable in the 'risk-adjustment' model, death episodes are only 16 percent more expensive than non-death episodes.*

*Outlier Episodes: As an alternative to excluding outlier episodes from the MSPB Measure, outlier episodes can instead be top-coded and/or bottom-coded. Rather than excluding episodes that are outliers, top-coding/bottom-coding assigns outliers the value of an episode at a specified threshold. Tables 6 through 10 in the appendix present the impacts of top-coding/bottom-coding episodes at the 99.9th/0.1th, 99.5th/0.5th, 99.0th/1.0th, 98.0th/2.0th, and 95.0th/5.0th percentiles, respectively, compared to a baseline that excludes outlier episodes at the 99th and*

*1st percentiles of the risk-adjusted episode cost distribution. When top-coded/bottom-coded at the 99.9th/0.1th, 99.5th/0.5th, and 99.0th/1.0th percentiles, at least 85 percent of MSPB Measure values change less than 3 percent. However, when top-coded/bottom-coded at the 98.0th/2.0th, and 95.0th/5.0th percentiles, at least 95% of MSPB Measure values change less than 3 percent (see Table 11).*

*Discharged AMA: Not only do episodes with an AMA discharge code make up a small percent of MSPB episodes (0.7%), AMA episodes have lower risk-adjusted spending than non-AMA episodes. ($13,851 vs. $19,025 for non-AMA). About 99% of hospitals experienced a change in their MSPB Measure values less than one percentage point when excluding AMA episodes (see Table 12).*

*Dual-Eligibles: 30% of episodes are flagged as dual-eligible beneficiaries; 18% of hospitals assigned an MSPB Measure have a beneficiary population consisting of at least 50% dual-eligible beneficiaries. Dual-eligible beneficiaries have $859 extra spending per episode than non-dual-eligible beneficiaries. If dual eligible are excluded, 43% of hospitals experience a change in their MSPB value of more than 1 percentage point (Table 13); including dual eligible in the risk adjustment model increases the R2 of the model by less than 0.001 and causes 12% of hospitals to change their MSPB Measure by more than 1 percentage point (Table 14).*

**2b3.3. Interpretation**

**What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

1. *Transfer Episodes:* Because transfer episodes are more inefficient than non-transfer episodes, regardless of the type of hospital (urban or rural), there are two main problems with including transfer episodes. First, because the observed cost relative to the predicted cost is high for transfer episodes (partly due to partial or full payments for two inpatient stays), including transfer episodes in the MSPB-Hospital measure may likely increase the MSPB-Hospital measure score of those hospitals most often engaging in transfers. These hospitals may not always have the capacity to handle these cases, and CMS may have an interest in ensuring medically appropriate transfers occur. Second, excluding transfer episodes addresses stakeholder concerns that neither the admitting nor receiving hospital is fully able to coordinate care. Stakeholders find it inappropriate to hold the transferring hospital responsible for services rendered by the receiving hospital, and it also may not be appropriate to hold the receiving hospital responsible for issues that arose prior to admission of a transferred patient. As a result, transfer episodes are excluded from the MSPB-Hospital measure calculation.

2. *Death Episodes:* Cases where the beneficiary dies during the episode are not eligible to be included in the MSPB-Hospital measure. Though the difference between cost for death and non-death episodes is relatively small compared to other exclusions, there are a few explanations for the exclusion of death episodes. First, including death episodes in MSPB-

Hospital measure calculation may create problematic incentives. Death episodes appear more efficient than non-death episodes; unlike non-death episodes, which have a slightly greater observed cost than expected cost, the observed cost for death episodes is much less than the expected cost. This is because beneficiaries with death episodes likely have shorter episodes (and therefore fewer services) than beneficiaries with non-death episodes with the same DRG. Because of this, including death episodes in MSPB-Hospital measure calculation may incentive low-quality care, as increased mortality rates could potentially improve hospitals' MSPB-Hospital measure scores by including episodes that appear more efficient. Second, episodes during which a beneficiary dies are "truncated;" in other words, costs that might have occurred if the beneficiary had not died are not observed due to death. Death episodes are incomplete episodes where significant data could be missing when death occurs early in the episode. To avoid including episodes of care with incomplete costs and problematic incentives, episodes during which a beneficiary dies are excluded from the MSPB-Hospital measure calculation.

3. *Outlier Episodes:* Outliers are excluded from the MSPB-Hospital measure calculation to avoid cases where a handful of high-cost and low-cost outliers have a disproportionate effect on each hospital's MSPB-Hospital measure score. While the correlation between the measure when excluding outliers versus when including outliers is extremely high (0.93), outlier episodes impact a small percentage of hospitals' MSPB-Hospital measure scores in a large and important way, as demonstrated by the differences in scores described in Appendix Table 2b3-5. The distribution of hospital risk-adjusted episode spending is significantly right-skewed: the 99th percentile is 3.6 times the value of the median, while the 1st percentile is less than half the value of the median. Excluding outliers based on risk-adjusted cost eliminates the episodes that deviate most from the spending levels one would have expected based on patient demographics and severity of illness.

4. *Overlapping episodes:* Episodes that begin during a prior episode's 30-day post-discharge period are excluded from MSPB-Hospital measure calculation. The impact of the exclusion on hospitals' MSPB-Hospital measure scores is minimal, and the correlation of the MSPB-Hospital measure calculated with and without implementing the overlapping episodes exclusion is high.

***<u>Previous response:</u>***

*Medicare Advantage or Part A Only: Due to missing claims problems, only beneficiaries enrolled in Medicare Parts A and B Fee-for-service are included in the sample.*

*Transfers: Adding transfers to the MSPB measure would significantly change hospital MSPB scores and make episode attribution more complicated. Assigning transfer episodes to the transferring hospital would avoid giving providers an incentive to transfer high-cost patients to game the system; however, once the transferring hospital transfers the patient, they may have little opportunity to coordinate or affect the patient's post-discharge care. Small rural hospitals, for example, often transfer patients in cases where they do not have the capacity to treat the patient within their current facilities. Assigning transfer episodes to the receiving hospital,* *however, incentivizes the initial hospital to transfer complex patients to improve their MSPB score. Further, post-acute care coordination may be difficult if the receiving hospital is out of area.*[16] *Public comment in the FY 2012 IPPS notice of proposed rulemaking voiced concern over attribution in transfer cases. In response, CMS excluded these types of transfers from the finalized MSPB Measure (76 FR 51621).*

*Death Episodes: In the baseline specification, cases where the beneficiary dies during the episode are not eligible to be included in the MSPB Measure. Episodes during which a beneficiary dies are "truncated"; in other words, costs that might have occurred if the beneficiary had not died are not observed due to death. To avoid including episodes of care with incomplete costs, episodes during which a beneficiary dies are excluded from the MSPB Measure calculation. As shown in 2b3.3., these episodes are typically high cost. In fact, the Dartmouth Atlas also notes that patients with chronic illness in their last two years of life account for about 32% of total Medicare spending, much of it going toward physician and hospital fees associated with repeated hospitalizations.*[17] *This evidence indicates that including death as a risk adjuster reduces the disparity in death/non-death episode cost. However, if death is a risk adjuster, hospitals could improve their MSPB score by increasing mortality rates. Further, using death as a risk adjuster implies that the risk adjustment model is no longer prospective, since events that occur during an episode now influence the model's expected cost.*

*Outlier Episodes: Outliers are excluded from the MSPB Measure calculation to avoid cases where a handful of high-cost and low-cost outliers have a disproportionate effect on each hospital's MSPB Measure score. The distribution of hospital risk-adjusted episode spending is significantly right-skewed: the 99th percentile is almost 4.5 times the value of the median, while the 1st percentile is only approximately 1/2 the value of the median. Excluding outliers based on risk-adjusted cost eliminates the episodes that deviate most from the spending levels one would expect based on patient demographics and severity of illness. Outliers are identified across all episodes rather than within a hospital; thus, some hospitals may have no outlier episodes excluded and others many have many.*

*Discharged AMA: Episodes with AMA index admissions should be eligible to be considered as index admissions, as the effect of excluding AMA episodes from the MSPB Measure calculation is minimal (as shown in Table 12). Additionally, episodes with an AMA discharge code make up a small percent of MSPB episodes, and AMA episodes on average have lower risk-adjusted spending than non-AMA episodes.*

*Dual-Eligibles: Medicare beneficiaries who are dually-eligible for Medicare and Medicaid are not excluded from the MSPB Measure to be consistent with NQF's position on not adjusting for potential demographic (sex or race) or socioeconomic factors.*

[16] As an alternative to completely assigning transfer episodes to either the transferring hospital or the receiving hospital, transfer episode costs could be split between both hospitals. A simple 50/50 weighting scheme would be one potential solution. To implement a 50/50 weighting scheme, each hospital receives 50% of the observed cost in the MSPB Amount numerator and 50% of the expected in the denominator of the MSPB Amount risk-adjustment factor ($\alpha_j$). This weighting scheme, however, does not take into account the length of stay at each hospital or the fact that the receiving hospital is in control of post-discharge spending. More complicated alternative weighting schemes (e.g., assigning a fixed weight to the receiving hospital and splitting the remaining weight based on the relative number of days the patient spends at each hospital) could be tailored to the particular application of the MSPB Measure, but these approaches would also increase the complexity of the MSPB Measure methodology.

[17] http://www.dartmouthatlas.org/keyissues/issue.aspx?con=2944

## 2B4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

### 2b4.1. Method

**What method of controlling for differences in case mix is used?**

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with** 854 **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

### 2b4.1.1 Risk Model Specifications

**If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

As described in Section S.9.3 on the Submission form, the MSPB-Hospital risk adjustment model broadly follows the CMS-HCC risk adjustment methodology, which is derived from Medicare Part A and B claims and is used in the Medicare Advantage (MA) program.[18] Although the MA risk adjustment model includes 24 age/sex variables, the MSPB-Hospital methodology does not adjust for sex and only includes 12 age categorical variables. Severity of illness is measured using 79 hierarchical condition category (HCC) indicators derived from the beneficiary's claims during the period 90 days prior to the start of the episode, the beneficiary's age, disability-status, and end-stage renal disease (ESRD) status, as well as an indicator of whether the beneficiary recently required long-term care, and the MS-DRG of the index hospitalization. The 79 HCC indicators are specified in Version 22 of the HCC model, and the HCC V22 model includes a mapping of ICD-9 diagnosis codes to CCs and ICD-10 diagnosis codes to CCs. The MSPB-Hospital risk adjustment methodology also includes status indicator variables for whether the beneficiary qualifies for Medicare through disability, age or ESRD. In addition, the model accounts for disease interactions by including interactions between HCCs and/or enrollment status variables that are included in the MA model. This is included because the presence of certain comorbidities increases costs in a greater way than predicted by the HCC indicators alone.[19] The MSPB-Hospital risk adjustment method does not control for the beneficiary's sex and race.

Just like the CMS-HCC model, the MSPB-Hospital risk adjustment approach uses an ordinary least squares (OLS) linear regression model. A separate OLS regression for predicted episode cost is calculated for each Major Diagnostic Category (MDC) that is determined by the MS-DRG of the index hospital stay. There are 26 different MDCs used in the risk adjustment model.

Severity of illness HCC indicators are created based on Medicare Part A and Medicare Part B diagnosis code information during the time 90 days prior to the start of an episode (i.e., 93 days

[18] Centers for Medicare and Medicaid Services, Office of the Actuary. "Announcement of Calendar Year (CY) 2014 Medicare Advantage Capitation Rates and Medicare Advantage and Part D Payment Policies and Final Call Letter." April 2013. https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/Announcement2014.pdf

[19] Centers for Medicare and Medicaid Services. "Medicare Managed Care Manual, Chapter 7 – Risk Adjustment, Section 70.2.7 – Disease and Disabled Interactions." 2014. https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/mc86c07.pdf

prior to the date of the index admission). Patients without a full 90-day look-back period (i.e., the beneficiary is not enrolled in both Medicare Part A and Medicare Part B for the 90 days prior to the episode) have their episodes excluded from the MSPB-Hospital measure. This 90-day period prior to the start of an episode is used to measure beneficiary health status; this look-back period ensures that each beneficiary's claims record contains sufficient fee-for-service data both for measuring spending levels and for risk adjustment purposes. As the length of the look-back period increases, there is a trade-off between the number of comorbidities captured and the number of false positives (i.e., diagnoses captured that may have been resolved). A longer look-back period, for example, will capture more comorbidities, while a shorter look-back period will capture fewer false positives. A longer look-back period will also decrease the number of episodes eligible to be included in the MSPB-Hospital measure calculation, since a beneficiary would be required to have a longer continuous stretch of pre-admission Medicare FFS enrollment to be included in the measure.

The MSPB-Hospital risk adjustment methodology also includes status indicator variables for whether the beneficiary qualifies for Medicare through Disability or End-Stage Renal Disease (ESRD); one can view these enrollment status variables as two additional severity of illness measures, however, these variables are generated from enrollment rather than diagnosis information.

Patients who reside in long-term care facilities typically require more intensive care—particularly more intensive post-acute care—than beneficiaries who live in the community, even holding constant illness severity measures. Thus, the risk adjustment method also includes an indicator of whether a beneficiary resides in a long-term care facility within the 90 days before the start of the episode as a non-diagnostic measure of severity of illness.

This measure assumes that the reason the patient is admitted to the hospital is largely outside the control of the hospital; thus, the risk adjustment measure also includes MS-DRG indicator variables as well. Additionally, the reason for admission directly affects payments and is predictive of post-acute care.

The relationship between comorbidities and episode cost may be non-linear in some cases. For instance, the marginal expected episode cost from having diabetes and congestive heart failure (CHF) may not be equal to the sum of the marginal expected cost from having diabetes and the marginal expected cost from having CHF. To account for these non-linearities, the MSPB-Hospital risk adjustment model also incorporates a series of interaction terms between HCCs and/or enrollment status variables that are included in the MA model.

The final set of explanatory variables in the risk adjustment model can be found in the "MSPB-Hospital Measure Information Form" available at the measure-specific web page URL identified in S.1 (see S.9.4.).

For reference, Appendix Table 2b4-A includes regression coefficients and standard errors for each of the covariates used in the risk adjustment models, stratified by MDC.

**2b4.2. Rationale if Not Risk Adjusted or Stratified**

**If an outcome or resource use component measure is <u>not risk adjusted or stratified</u>, provide <u>rationale and analyses</u> to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

N/A

**2b4.3. Conceptual/Clinical and Statistical Methods and Criteria**

**Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

The CMS-HCC model was selected based on previous studies evaluating its appropriateness for use in risk adjusting Medicare claims data. This model was developed specifically for use in the Medicare population, meaning that it accounts for conditions found in the Medicare population and is calibrated on Medicare Fee-for-Service (FFS) beneficiaries. In addition, the CMS-HCC model is annually updated for changes in coding practices (e.g., the transition from ICD-9 to ICD-10 codes) and is exhaustive on these code sets. Because the CMS-HCC model has already been extensively tested, we focus on adapting the CMS-HCC model to the MSPB-Hospital measure methodology.[20]

A number of studies have shown that socioeconomic status is associated with the amount of resources used during the period in which patients are hospitalized as well as during post-acute care. A larger proportion of low-income Medicare beneficiaries tended to use inpatient services in a given year compared to patients with higher incomes (25% and 17%, respectively). Lower-income beneficiaries are also twice as likely to use home health services as Medicare beneficiaries earning higher incomes.[21] End-of-life care for Medicare beneficiaries who are Black or Hispanic is substantially different than the end-of-life hospital services that Medicare beneficiaries who are White receive. Much of the variation in end-of-life care is due to differences in utilization levels among hospitalized patients. Beneficiaries who are Black and who are Hispanic are significantly more likely to be admitted to the ICU than beneficiaries who are White, and minorities also receive significantly more intensive procedures, such as resuscitation and cardiac convers, mechanical ventilation, and gastrostomy for artificial nutrition.[22]

According to a 2014 National Quality Forum report, the mechanisms underlying differences in resource use by socioeconomic status and race are complex and may be impacted by factors such as financial resources, community resources, historical and current discrimination, and reduced access to preventive services. Provider assumptions or implicit biases may impact quality of care for beneficiaries of different races. These factors may result in inefficient care, increased disease severity, or greater morbidity,[23] leading to higher Medicare spending for beneficiaries depending on socioeconomic status or race.

Given the conceptual and empirical relationship between income, race, and resource use, we analyzed both socioeconomic status (SES) and sociodemographic status (SDS), where SDS is

[20] Pope, Gregory C., John Kautter, Melvin J. Ingber, Sara Freeman, Rishi Sekar, and Cordon Newhart. “Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report.” RTI International: March 2011.

[21] Kaiser Family Foundation. “Medicare Chartbook” Fourth Edition, 2010. http://www.kff.org/medicare/upload/8103.pdf

[22] Hanchate, Amresh, et al. “Racial and Ethnic Differences in End-of-Life Costs: Why do Minorities Cost More than Whites?” Archives of Internal Medicine. 2009; 169(5):493-504.

[23] National Quality Forum. “Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors: Technical Report.” National Quality Forum: August 2014.

defined as SES and race considered together. To determine SES, we used the United States Census Bureau's 2014 American Community Survey (ACS) 5-year estimates. The ACS dataset "Ratio of Income to Poverty Level of Families in the Past 12 Months" contains variables that provide population estimates of ranges of income-to-poverty ratios by ZIP code. Because individual family members may pool financial resources to provide care for older relatives, we used family income-to-poverty ratio in SES analysis instead of individual income-to-poverty ratio to better represent household decisions.[24] For a given ZIP code, the family income-to-poverty ratio dataset contains the variables: "Under .50", ".50 to .74", ".75 to .99", "1.00 to 1.24", "1.25 to 1.49", "1.50 to 1.74", "1.75 to 1.84", "1.85 to 1.99", "2.00 to 2.99", "3.00 to 3.99", "4.00 to 4.99", and "5.00 and over". Each of these variables gives the count of families in a given ZIP code whose income falls into that category range of income-to-poverty level. To illustrate, if the value for the ".50 to .74" variable is 10,000 for a particular ZIP code, that means that 10,000 families in that ZIP code have incomes that are between 50% and 74% of the federal poverty threshold.

The Enrollment Database (EDB) provided data on beneficiary race, and we look at race because race tracks with SES, and we wanted to see the impact on hospitals' performance on the MSPB-Hospital measure. While the EDB provides data on all race categories, there are concerns with the validity of the race categories other than Black and White (e.g., Asian, Hispanic, North American Native) due to underreporting in those categories.[25] As a result, we categorized beneficiaries as Black or Non-Black, where Non-Black is defined as all other race categories. The EDB also provided the ZIP codes for beneficiaries included in the sample. We then linked these beneficiary ZIP codes to the ACS ZIP code-level data on family income-to-poverty ratio to estimate the income-to-poverty ratio for each beneficiary with an MSPB-Hospital episode.

Using these data, we conducted a number of analyses related to disparities by population group. For race categories, we produced an estimated distribution of beneficiaries by income ratio (see Section 2b4.4b. for analysis). Additionally, we sought to determine the effect of incorporating SES or SDS into our risk adjustment model by determining the difference in MSPB-Hospital measure scores when including SES or SDS. We also analyzed correlation between MSPB-Hospital measure scores calculated with and without SES or SDS. The outcome of these analyses is discussed in Section 2b4.5.

***Previous response:***

*To account for case-mix variation and other factors, the MSPB risk-adjustment methodology broadly follows the CMS-HCC risk-adjustment methodology, which CMS uses to estimate Medicare Advantage (MA) premium adjustments.[26] Medicare also uses the HCC model to risk-adjust spending in: the Shared Savings Program Accountable Care Organizations (implemented in 2012) and the Medicare Physician Quality and Resource Use Reports (implemented in 2009). The accuracy of the ICD-9 codes used to create HCCs has also been evaluated in previous studies, and all studies found high positive predictive values for Medicare claims-based diagnosis of*

[24] Deaton, Angus S. and Paxson, Christina. *Chapter 6: Measure Poverty among the Elderly.* (Inquiries in the Economics of Aging, University of Chicago Press, January 1998), 171. https://core.ac.uk/download/pdf/6870973.pdf

[25] Zaslavsky, Alan M, John Z Ayanian, and Lawrence B Zaborski. "The Validity of Race and Ethnicity in Enrollment Data for Medicare Beneficiaries." *Health Services Research* 47.3 Pt 2 (2012): 1300–1321. PMC. Web. 28 Oct. 2016.

[26] Centers for Medicare and Medicaid Services, Office of the Actuary. "Announcement of Calendar Year (CY) 2009 Medicare Advantage Capitation Rates and Medicare Advantage and Part D Payment Policies." April 2008. http://www.cms.gov/MedicareAdvtgSpecRateStats/Downloads/Announcement2009.pdf

*acute myocardial infarction (AMI), chronic kidney disease (CKD), heart failure, coronary artery disease, diabetes, hypertension, and stroke with a diagnosis based on structured hospital record review.*[27,28,29] *A 2003 study found that CMS "administrative data was found to have diagnoses and conditions that were highly specific but that vary greatly by condition in terms of sensitivity."*

*Severity of illness is measured using 70 HCC indicators derived from the beneficiary's claims during the period 90 days prior to the start of the episode, an indicator of whether the beneficiary recently required long-term care, as well as the MS-DRG of the index hospitalization. The MSPB risk-adjustment methodology also includes status indicator variables for whether the beneficiary qualifies for Medicare through Disability or End-Stage Renal Disease (ESRD) and whether a beneficiary resides in a long-term care facility. Because the relationship between comorbidities' episode cost may be non-linear, the model includes interactions between HCCs and/or enrollment status variables. The MSPB risk-adjustment method does not control for the beneficiary's sex and race, but does include 12 age categorical variables. For a complete list of MSPB risk-adjustment variables, see the "MSPB Measure Information Form" available on QualityNet at the link provided in S.1.*

*All explanatory variables are calculated during the 90 days prior to the start of an episode. Calculating all health status variables prior to the start of an episode avoids the endogeneity problem which could occur if the diagnosis codes a hospital uses are included in the risk-adjustment model. Using claims data during the episode would incentivize hospitals to inflate the number of co-morbidities (i.e., number of diagnosis codes) that a beneficiary has to make their health status appear worse.*

*The MSPB risk-adjustment methodology (along with the entire MSPB methodology) was also put through official notice and comment rulemaking. The majority of commenters supported the risk adjustment for age and severity of illness. Some suggested further adjustment for race, sex, or socioeconomic factors, but Acumen and CMS opted to maintain consistency with the NQF's position against adjusting for these factors.*

**2b4.4a. Results of Analyses to Select Risk Factors**

**What were the statistical results of the analyses used to select risk factors?**

The MSPB Measure broadly replicates the CMS-HCC model. The literature has extensively tested the use of the HCC model as applied to Medicare claims data. Although the variables in the HCC model were chosen to predict annual cost, CMS also uses this risk adjustment model in a number of other settings (e.g., ACOs and physician QRUR programs).

Recalling that the risk model relies on the existing CMS-HCC model, more information on factors included in the CMS-HCC model can be found at Pope et al. 2011.[30]

***Previous response:***

[27] Kiyota, Uka, et al. "Accuracy of Medicare Claims-Based Diagnosis of Acute Myocardial Infarction: Estimating Positive Predictive Value on the Basis of Review of Hospital Records." American Heart Journal. 148(1): 99-104, July 2004.

[28] Winkelmayer, W. C., et al. "Identification of Individuals with CKD from Medicare Claims Data: A Validation Study." Am J Kidney Dis. 46(2): 225-232, Aug 2005.

[29] Birman-Deych, Elena, et al. "Accuracy of ICD-9-CM Codes for Identifying Cardiovascular and Stroke Risk Factors." Medical Care. 43(5): 480-485, May 2005.

[30] Pope et al., "Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report."

*The MSPB Measure broadly replicates the CMS-HCC model. The literature has extensively tested the use of the HCC model as applied to Medicare claims data.*[31] *Although the variables in the HCC model were chosen to predict annual cost, CMS also uses this risk-adjustment model in a number of other settings (e.g., ACOs and physician QRUR programs).*[32]

**2b4.4b. Analyses and Interpretation on SDS Factors**

**Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)**

To examine how race tracks with SES, we produced an estimated distribution of beneficiaries by income-to-poverty ratio and by race. At the hospital level, there is a minimal impact on measure score. This is discussed further in Section 2b4.5 below. Among the lower income-to-poverty ratio ranges (i.e., below or near the poverty level), there was a greater percentage of beneficiaries who were Black (19%) when compared to the percentage of beneficiaries who were Non-Black (11%). Among higher income-to-poverty ratio ranges (i.e., ratio above 5), there were a greater percentage of beneficiaries who were Non-Black (31%) compared to the percentage of beneficiaries who were Black (22%). Appendix Table 2b4-3 details the breakdown of income-to-poverty ratio ranges by race category.

The outcome of analyses testing SES and SDS in risk adjustment is discussed in the following section, Section 2b4.5.

**2b4.5. Method**

**Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model <u>or</u> stratification approach** (*describe the steps—do not just name a method; what statistical analysis was used*)

This section discusses the methodology used to analyze the following aspects of risk adjustment: (i) specification of the look-back period and stratification options, (ii) validity of current risk adjustment model, and (iii) evaluation of including SES and SDS.

Empirical evaluations of (i) focused on two specifications: first, the look-back period used to calculate comorbidities, and second, the methodology used to stratify the risk adjustment models. For the look-back period, the two options were 90-days, which is the period used in the current measure calculation, and 1 year. For stratifying the risk adjustment model, the options were to use only MDC, which is the current specification, or to use a combination of MDC and institutional status (i.e., whether a beneficiary is in long term care as determined using MDS data).

To demonstrate the validity of the MSPB risk adjustment methodology, we calculated the distribution of episode spending and R-squared by decile to examine the model's ability to predict both very low and high cost episodes. Specifically, we created a "risk score" for each episode calculated as the predicted cost values from each episode divided by the national

[31] Pope, Gregory C., John Kautter, Melvin J. Ingber, Sara Freeman, Rishi Sekar, and Cordon Newhart. "Evaluation of the CMS-HCC Risk-Adjustment Model: Final Report." RTI International: March 2011.

[32] Department of Health and Human Services, Centers for Medicare and Medicaid Services, Medicare Program; Medicare Shared Savings Program: Accountable Care Organizations, Proposed Rule, Federal Register, April 7, 2011 76(67):19528–654.

average predicted cost value. After arranging episodes into deciles based on the risk score, we calculated the predictive ratio for each decile using the formula of average(expected cost)/average(observed cost) for all episodes in each decile. In addition, we calculated a "90/10 ratio," comparing the average cost of episodes in the first decile to the average cost of episodes in the tenth decile for observed costs and risk-adjusted costs. Risk-adjusted costs were calculated in two ways, by ratio and by residual. For the ratio calculation, we calculated risk adjusted cost for each episode as (observed cost/expected cost), multiplied by a national mean cost. For the residual calculation, we calculated risk adjusted cost for each episode as (observed cost – expected cost) + national mean observed cost.

We examined the impact of including SES or SDS into our risk adjustment model with three tests: F-test of significance, difference in MSPB-Hospital measure scores, and correlation between MSPB-Hospital measure scores. First, we performed F-tests to assess the significance of SES and SDS on predicting resource use. The F-test revealed many significant p-values at the MDC level (see Appendix Table 2b4-4 and 2b4-6). This indicates that SES and SDS are likely predictive factors for determining resource use among beneficiaries for the relevant MDCs.

Overall, SES and SDS are likely predictive of variation in resource use. However, when including SES or SDS in our risk adjustment regression with other variables, the very minor change in hospital scores indicates that SES and SDS effects on hospital scores are largely captured through existing risk adjustment variables. We sought to determine the effect of incorporating SES or SDS into our risk adjustment model by determining the difference in MSPB-Hospital measure scores when including SES or SDS. In both cases, the differences in MSPB-Hospital measure scores were minimal (see Appendix Table 2b4-5 and 2b4-7). When including SES in risk adjustment, the MSPB-Hospital measure score for 97% of hospitals changed by ±0.01 or less. When including SDS in risk adjustment, the MSPB-Hospital measure score for 95% of hospitals changed by ±0.01 or less. Finally, we analyzed the correlation between MSPB-Hospital measure scores calculated with and without SES or SDS. The MSPB-Hospital measure scores calculated with and without SES were highly correlated (>0.998), as were measure scores calculated with and without SDS (>0.997). Because inclusion of SES and SDS factors has a minimal impact on the measure score and due to the high correlation values, we do not believe that including SES or SDS factors in the MSPB-Hospital risk adjustment methodology is appropriate.

***Previous response:***

*Because the CMS-HCC model has already been extensively tested, we focus on adapting the CMS-HCC model to the MSPB Measure methodology. To empirical evaluate the MSPB risk-adjustment methodology, we analyzed two specifications of the modified CMS-HCC risk-adjustment methodology by using R2 to measure model ability to explain variation: (1) evaluate the health status variables in the risk-adjustment by using one year of data prior to calculate comorbidities rather than 90 days; and (2) evaluate options for stratifying the risk-adjustment model (e.g., by MDC, MDC/Institutional Status). To demonstrate the validity of the MSPB risk-adjustment methodology, we (3) calculated the distribution of episode spending and R-squared by decile to examine the model's ability to predict both very low and high cost episodes. Specifically, we created a "risk score" for each episode calculated as the predicted values from each episode divided by the national average predicted value. After arranging episodes into deciles based on the risk score, we calculated the R-squared for each decile using the formula 1-(SSE/SST), where SSE = the sum of (episode observed spending – episode predicted spending) and SST = the sum of (episode observed spending – average overall observed spending).*

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.*

***If stratified, skip to 2b4.9***

**2b4.6. Statistical Risk Model Discrimination Statistics**

**Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

The average R-squared for the MSPB-Hospital measure risk adjustment model across all MDCs is 0.3014. The overall R-squared, calculated by comparing residuals to the difference between observed costs and the national mean cost across all MDCs, is 0.4757. Appendix Table 2b4-A also includes regression coefficients and standard errors for each of the covariates used in the risk adjustment models. More information on discrimination testing for the CMS-HCC model can be found at Pope et al. 2011.[33]

***Previous response:***

*The overall R-squared for the MSPB Measure risk adjustment model described in S.9.2. through S.9.4. is 0.4621. For your reference, the "Additional Information" Appendix beginning on page 24 of the "Scientific Acceptability" section also includes regression coefficients, standard error, and p-values of the covariates used in the risk-adjustment models. Recalling that the risk model relies on the existing CMS-HCC model, more information on discrimination testing for the CMS-HCC model can be found at Pope et al. 2011.*[34]

**2b4.7. Statistical Risk Model Calibration Statistics**

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

1. *Evaluate options for look-back periods*: When changing the HCC "look-back" period from 90 days to 365 days: (i) 6.7% of episodes are dropped and (ii) the overall model fit (i.e., average of R-squared across all MDCs) decreases from 0.3014 to 0.2997. The R-squared, when calculated overall across MDCs, decreases from 0.4757 to 0.4736. More detailed statistics are shown in Appendix Table 2b4-1.

2. *Evaluate options for stratification of risk adjustment model*: When stratifying the risk adjustment model by MDC only, but with an indicator for institutional status (e.g., Long-Term Institutional (LTI) indicator) (current specification), the average R-squared across MDCs is 0.3014 and the overall R-squared is 0.4757. On the other hand, when stratifying the risk adjustment model by MDC, but with separate regressions for institutional and community beneficiaries, the average R-squared across MDCs is 0.3060 and the overall R-squared is 0.4778. In addition, when averaging across MDCs, 60.27% of regression variables have a p-value of less than 0.1 when using the MDC model, while only 48.93% of regression variables have a p-value of less than 0.1 when using the MDC/Institutional model. Further statistics by MDC are shown in Appendix Table 2b4-2.

***Previous response:***

[33] Ibid.

[34] Ibid.

*1. Assessing the use of one year of data prior to the index admission to calculate comorbidities in the risk adjustment methodology rather than 90 days: When changing the HCC "look-back" period from 90 days to 365 days: (i) 6% of episodes are dropped (see Table 19 in the appendix) and (ii) the model fit (i.e., R-squared) decreases from 0.4621 to 0.4601. The impact analysis also reveals that, despite the drop in episodes included and a decrease in model fit, most hospitals experience only a small change in their MSPB Measure values when switching the "look-back" period from 90 days to 365 days; in fact, Table 20 in the appendix shows that 78% of hospitals experience a gain or loss in the MSPB Measure values of less than 1 percentage point.*

*2. Evaluating options for stratifying the risk adjustment model (e.g., by MDC, MDC/Institutional Status): When stratifying the risk-adjustment model by MDC with a Long-Term Institutional (LTI) indicator (current specification), the R-squared is 0.4621. On the other hand, when stratifying the risk-adjustment model by MDC, but with separate regressions for institutional and community beneficiaries, the R-squared is 0.4645. When stratifying the risk-adjustment model by MDC, but with separate regressions for MDC type (i.e., MED, SURG), the R-squared is 0.4636. The MDC option was preferred because: (i) the improvement in R-squared is very small when moving to the MDC/Institutional Status specification and (ii) increasing the number of stratifications increases the risk of over-fitting, especially for MDCs with relatively few admissions.*

**2b4.8. Statistical Risk Model Calibration**

**Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

3. *Evaluate the validity of the risk adjustment model:* Table 1 below shows predictive ratios by risk decile for the MSPB-Hospital measure. The table shows that the model has consistent predictive ratios across risk score deciles, with the first decile having a predictive ratio of 0.994 and the tenth decile having a predictive ratio of 1.011.

**Table 1: Predictive Ratios by Risk Decile for MSPB-Hospital**

| Decile | Number of Episodes | Average Observed Standardized Spending | Average Expected Standardized Spending | Predictive Ratio |
|---|---|---|---|---|
| 1 | 542,061 | $8,570.70 | $8,621.74 | 0.994080 |
| 2 | 542,073 | $11,166.26 | $11,288.23 | 0.989192 |
| 3 | 542,060 | $13,134.89 | $13,136.85 | 0.999850 |
| 4 | 542,059 | $15,066.59 | $14,970.63 | 1.006413 |
| 5 | 542,063 | $17,257.92 | $17,122.40 | 1.007915 |
| 6 | 542,064 | $19,242.71 | $19,377.29 | 0.993055 |
| 7 | 542,064 | $21,411.22 | $21,642.11 | 0.989332 |
| 8 | 542,053 | $24,151.26 | $24,304.96 | 0.993676 |
| 9 | 542,072 | $28,864.21 | $28,920.72 | 0.998046 |
| 10 | 542,064 | $46,105.10 | $45,585.95 | 1.011388 |

The 90/10 ratio calculation shows that the risk adjustment model does effectively shrink the dispersion of the cost distribution. At the observed cost level, the 90/10 ratio is 6.22. The costs risk-adjusted by ratio have a 90/10 ratio of 3.40, and the costs risk-adjusted by residual have a 90/10 ratio of 3.21.

***Previous response:***

*3. Calculate the distribution of episode spending and R-squared by decile to show that the MSPB risk adjustment methodology does equally well predicting spending through all values of the model: The R-squared in the 3rd through 9th deciles are lower than overall R-squared in Table A below (includes outlier episodes) as well as Table B below (excludes outlier episodes). The R-squared in the 6th and 7th deciles are relatively low, ranging from approximately 1% to 3%. Additionally, the R-squared is always higher in Table B when outlier episodes are excluded.*

***Table A: Distribution of Spending and R-Squared by Decile*[*]* (Includes Outlier Episodes)***

| Decile | Episode Count | Min Risk Score | Max Risk Score | Avg. Obs Spending | Avg. Pred Spending** | Difference | R-Squared |
|---|---|---|---|---|---|---|---|
| 1 | 446,268 | -0.38 | 0.46 | $7,442 | $7,365 | $77 | 0.7774 |
| 2 | 446,234 | 0.46 | 0.56 | $9,607 | $9,763 | -$156 | 0.5861 |
| 3 | 446,197 | 0.56 | 0.65 | $11,472 | $11,506 | -$34 | 0.3876 |
| 4 | 446,234 | 0.65 | 0.74 | $13,379 | $13,276 | $103 | 0.2365 |
| 5 | 446,260 | 0.74 | 0.85 | $15,164 | $15,114 | $50 | 0.1194 |
| 6 | 446,205 | 0.85 | 0.98 | $17,452 | $17,350 | $101 | 0.0229 |
| 7 | 446,512 | 0.98 | 1.14 | $20,047 | $20,226 | -$179 | 0.0100 |
| 8 | 445,951 | 1.14 | 1.31 | $23,108 | $23,237 | -$128 | 0.0858 |
| 9 | 446,130 | 1.31 | 1.66 | $27,830 | $27,631 | $199 | 0.1680 |
| 10 | 446,339 | 1.66 | 20.09 | $45,115 | $45,148 | -$33 | 0.6903 |
| TOTAL | 4,462,330 | -0.38 | 20.09 | $19,062 | $19,062 | $0 | 0.4621 |

*Note: *Decile are based on risk score calculated as ratio of predicted spending over national average predicted spending.*
***Predicted spending is the predicted value from the regression.*

***Table B: Distribution of Spending and R-Squared by Decile*[*]* (Excludes Outlier Episodes)***

| Decile | Episode Count | Min Risk Score | Max Risk Score | Avg. Obs Spending | Avg. Pred Spending** | Difference | R-Squared |
|---|---|---|---|---|---|---|---|
| 1 | 437,305 | 0.04 | 0.46 | $7,087 | $7,348 | -$262 | 0.8644 |
| 2 | 437,313 | 0.46 | 0.56 | $9,140 | $9,730 | -$590 | 0.6989 |
| 3 | 437,309 | 0.56 | 0.65 | $10,905 | $11,458 | -$553 | 0.5135 |
| 4 | 437,248 | 0.65 | 0.74 | $12,776 | $13,213 | -$436 | 0.3249 |
| 5 | 437,370 | 0.74 | 0.84 | $14,596 | $15,035 | -$439 | 0.1744 |
| 6 | 437,310 | 0.84 | 0.98 | $16,887 | $17,247 | -$360 | 0.0329 |
| 7 | 437,298 | 0.98 | 1.14 | $19,566 | $20,124 | -$558 | 0.0140 |
| 8 | 437,320 | 1.14 | 1.31 | $22,534 | $23,144 | -$609 | 0.1288 |
| 9 | 436,500 | 1.31 | 1.66 | $27,237 | $27,502 | -$265 | 0.3627 |
| 10 | 438,118 | 1.66 | 20.17 | $44,304 | $45,039 | -$735 | 0.7752 |
| TOTAL | 4,373,091 | 0.04 | 20.17 | $18,506 | $18,987 | -$481 | 0.5978 |

*Note: [*]Deciles are based on risk score calculated as ratio of predicted spending over national average predicted spending.*
*[**]Predicted spending is the Winsorized and renormalized predicted value.*

**2b4.9. Results of Risk Stratification Analysis:**

N/A

**2b4.10. Interpretation**

**What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i.*e., what do the results mean and what are the norms for the test conducted*)

The R-squared values for the model, which measure the percentage of variation in results predicted by the model, are in line with or are higher than the values presented in similar analyses of risk adjustment models.[35]

1. *Evaluate options for look-back periods*: As both the model fit and number of episodes included decrease when moving to a 365 day window for calculating comorbidities, the MSPB-Hospital risk adjustment model appropriately uses a 90 day period.

2. *Evaluate options for stratification of risk adjustment model*: These numbers justify the continued use of stratifying by MDC because: (i) the improvement in R-squared is very small when moving to the MDC/Institutional Status specification, (ii) increasing the number of stratifications by including institutional status increases the risk of over-fitting, especially for MDCs with relatively few admissions, and (iii) more variables are statistically significant predictors in the MDC model as determined by a p-value of less than 0.1, which is generally accepted as statistically significant.

3. *Evaluate the validity of the risk adjustment model:* The risk decile table shows that the risk adjustment model has consistent predicted spending for all deciles. Predictive ratios close to 1 indicate that expected spending is accurately predicting observed spending. The maximum variation from 1 is in the tenth decile, with a predictive ratio of 1.011. Overall, this table shows that the model is accurately predicting observed spending, regardless of decile.

A larger 90/10 ratio shows that the distribution of costs has a wider spread. This is an effective measure of dispersion, as compared to the standard deviation, because episode costs are skewed towards high-cost outliers. The 90/10 ratio, dropping by 45% and 48% for the ratio and residual calculations, respectively, does show that the risk adjustment for the MSPB-Hospital measure effectively reduces the dispersion in episode spending. Other investigations of the 90/10 ratio have found reductions of dispersion ranging from 20% to 48%.[36] This shows that the risk adjustment model does account for high-cost episodes and controls for the variance in observed spending.

[35] Ibid, 6.

[36] MaCurdy, Thomas et al. "Challenges in the Risk Adjustment of Episode Costs." CMS, February 2010. Available online at https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Reports/downloads/MaCurdy_ERA_2010.pdf.

***Previous response:***

*1. Assessing the use of one year of data prior to the index admission to calculate comorbidities in the risk adjustment methodology rather than 90 days: When the FFS continuous enrollment requirement starts from 365 days prior to the start of the episode instead of 90 days prior to the start of the episode, there is no trade-off between the number of episodes included in the MSPB Measure and the model fit. In fact, both the number of episodes included and the model fit decrease (i.e., get worse).*

*2. Evaluating options for stratifying the risk adjustment model (e.g., by MDC, MDC/Institutional Status): The R-squared between the different options for stratifying the risk-adjustment model are comparable, indicating that the output is not very different. However, when separate regressions for the community/institutional model or the MED/SURG MDC model are run, degrees of freedom are lost and may cause over-fitting of the model.*

*3. Calculate the distribution of episode spending and R-squared by decile to show that the MSPB risk adjustment methodology does equally well predicting spending through all values of the model: Based on the distribution of spending and R-squared by decile, we believe that the MSPB risk-adjustment methodology is robust and fit consistently across deciles.*

### 2b4.11. Optional Additional Testing for Risk Adjustment

(*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

N/A

***Previous response:***

*Limited additional testing was performed because the MSPB Measure risk-adjustment methodology is intended to closely follow the established and extensively tested CMS-HCC risk-adjustment methodology. As previously discussed, however, we did test stratifying the model by MDC/Institutional Status rather than just stratifying the model by MDC. We also tested different look-back periods from the current 90 days.*

## 2B5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

### 2b5.1. Method

**Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Our method to determine clinically meaningful differences in MSPB-Hospital measure scores consists of stratifying MSPB-Hospital measure scores by meaningful hospital characteristics, and comparing those results to expected findings discussed in the literature. Stratification is performed for each of the following characteristics: urban/rural location and hospital size;

urban/rural location and geographic region;[37] and teaching status. We analyze the distribution of MSPB-Hospital measure scores for subgroups defined by these characteristics, as well as for the overall population. The purpose of this analysis is to ensure that MSPB-Hospital measure scores vary in a manner consistent with expectations. That is: the literature has identified certain characteristics with a meaningful relationship to hospital performance, and this analysis stratifies MSPB-Hospital measure scores by those same characteristics. This analysis is therefore slightly different than the reliability and validity analyses discussed in Sections 2a2 and 2b2, since it specifically seeks to confirm that the MSPB-Hospital measure behaves as expected with respect to well-documented and meaningful hospital characteristics.

***Previous response:***

*MSPB summary statistics include the percentile distribution of the MSPB score both overall and by hospital type (e.g., urban/rural status, bed size, region, teaching status). Although poor MSPB scores could be due to low quality care, it could also be the case that unobservable factors (e.g., large populations of patients for whom English is a second language, low adherence to treatment regimens) outside of hospitals' control make these hospitals perform worse. To identify hospitals that treat a large number of socioeconomically disadvantaged patients, the following analysis also classifies hospitals by their Disproportionate Share Hospital (DSH) percentage.*[38]

**2b5.2. Results**

**What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Key findings include: (1) the highest single MSPB-Hospital measure score is more than five times higher than the hospital with the lowest MSPB-Hospital measure score; (2) the MSPB-Hospital measure score at the 90th percentile is almost 23 percent greater than the MSPB-Hospital score at the 10th percentile; (3) the average MSPB-Hospital measure score for rural hospitals is almost five percent lower than the average MSPB-Hospital measure score for urban hospitals; (4) the average MSPB-Hospital Measure score in the West South Central region is the highest for both urban and rural hospitals, followed by the Mid Atlantic and New England for urban hospitals and the East South Central and East North Central for rural hospitals; and (5) the average MSPB-Hospital measure score for teaching hospitals is higher than the measure score for non-teaching hospitals. Appendix Tables 2b5-1 through 2b5-4 present these results.

***Previous response:***

*Key findings include: (1) the hospital with the highest MSPB score costs Medicare more than six times as much as the lowest cost hospital; (2) hospitals at the 90th percentile MSPB Measure cost Medicare 25 percent more per episode than hospitals at the 10th percentile; (3) rural hospitals out-perform urban hospitals; (4) the average MSPB Measure value in New England and*

[37] The geographic regions used in this analysis are drawn from the census regions and divisions used by the U.S. Census Bureau. See "Census Regions and Divisions of the United States." U.S. Census Bureau. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

[38] The Medicare DSH patient percentage is equal to the sum of the percentage of Medicare inpatient days attributable to patients entitled to both Medicare Part A and Supplemental Security Income and the percentage of total inpatient days attributable to patients eligible for Medicaid but not eligible for Medicare Part A.

*the West South Central regions are the highest for both urban and rural hospitals; (5) teaching hospitals have higher average spending levels, but they also have higher expected spending amounts (due to a sicker patient case mix); and (6) hospitals with a large number of DSH-eligible patients are not significantly less efficient than hospitals with few DSH beneficiaries. Tables 15 through 18 in the appendix present these results.*

### 2b5.3. Interpretation

**What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

There exists clinically/practically significant variation in MSPB-Hospital measure scores, which indicates the measure's ability to capture differences in performance. There also exists significant variation in MSPB-Hospital measure scores when considered in light of certain clinically meaningful hospital characteristics. As noted above, rural hospitals tend to have lower MSPB-Hospital measure scores than urban hospitals, and the West South Central region has the highest average MSPB-Hospital measure score for both urban and rural hospitals. As mentioned in section S.11, low MSPB-Hospital measure score(s) indicates that the hospital or set of hospitals have low MSPB-Hospital amount(s) (i.e., risk-adjusted spending); measure scores less than 1 indicate that the MSPB-Hospital amount is less than the national episode-weighted median MSPB-Hospital amount across all hospitals during the given performance period. The results can be interpreted to mean that hospitals with lower MSPB-Hospital measure scores have lower risk-adjusted spending than other hospitals.

Our findings regarding variation in the MSPB-Hospital measure, particularly with respect to clinically meaningful hospital characteristics, are consistent with existing literature. Research by the Dartmouth Institute for Health Policy & Clinical Practice has found significant variation in hospital expenditures for the Medicare population,[39] which is consistent with our findings regarding significant variation across MSPB-Hospital measure score percentiles. Dartmouth has also found significant variation with respect to characteristics considered by our analysis. In particular, their research has found that southern and northeastern states generally have high Medicare utilization, and that certain urban areas had higher Medicare utilization.[40] These findings within the literature are consistent with our stratified findings of the MSPB-Hospital measure score by geographic region and urban/rural hospital. Other literature also found that academic centers tend to have higher Medicare spending, which is consistent with our findings about teaching hospitals. [41]

***Previous response:***

*There exists significant variation in spending relative to the typical hospital. For example, hospitals at the 90th percentile use 25 percent more resources per episode than hospitals at the 10th percentile. These figures also vary across hospital characteristics.*

[39] Fisher, Elliott et al. "Health Care Spending, Quality, and Outcomes." The Dartmouth Institute for Health Policy and Clinical Practice. February 27, 2009. http://www.dartmouthatlas.org/downloads/reports/Spending_Brief_022709.pdf

[40] Skinner, Jonathan et al. "A New Series of Medical Expenditure Measures by Hospital Referral Region: 2003-2008". The Dartmouth Institute for Health Policy and Clinical Practice. June 21, 2011. http://www.dartmouthatlas.org/downloads/reports/PA_Spending_Report_0611.pdf

[41] Romley, John et al. "Spending and Mortality in US Acute Care Hospitals." Am J Manag Care. 2013;19(2):e46-e54

## 2B6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

***If only one set of specifications, this section can be skipped*.**

N/A

## 2B7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

### 2b7.1. Method

***Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data*** *(or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (describe the steps—do not just name a method; what statistical analysis was used)*

Since CMS uses Medicare claims data to calculate the MSPB-Hospital measure, the required data are readily available and retrievable without undue burden. In fact, Acumen has already acquired all the data needed and has already calculated the MSPB-Hospital measure.

### 2b7.2. Results

***What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?*** *(e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each)*

N/A

### 2b7.3. Interpretation

***What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias?*** *(i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)*

N/A

## Feasibility

F.1. **Byproduct of Care Processes**

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

F.1.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

F.2. **Electronic Sources**

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

F.2.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*)

ALL data elements are in defined fields in electronic claims

F.2.1a. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.

F.2.2. **If this is an eMeasure,** provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.

**Attachment:**

F.3. **Data Collection Strategy**

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

F.3.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

CMS uses Medicare administrative claims data that hospitals submit to CMS for payment to calculate the MSPB-Hospital measure. As a result, the required data are readily available and retrievable without undue burden. These claims data used are maintained by CMS's OIS. These data undergo additional quality assurance checks during measure development and maintenance. Specifically, CMS has in place several hospital auditing programs used to assess overall claims code accuracy, ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analyses to identify potential problem areas and detect fraud. CMS also audits important data fields, including diagnosis and procedure codes, as well as other elements that are consequential to payment. Specifically, CMS works with Program Safeguard Contractors (PSCs)/Zone Program Integrity Contractors (ZIPCs) to ensure program integrity; the agency also uses Comprehensive Error Rate Testing (CERT) Contractors to ensure that Medicare payments are correct. Between 2005 and 2015, CERT estimates that proper payment, which is payments that met Medicare coverage, coding, and billing rules, ranged from 87.3 to 96.4 percent of total payments each year.[1] CMS continues to perform successful corrective actions and give providers additional education to ensure accurate billing. To ensure claims completeness and inclusion of any corrections, the measure is calculated using data with a 3 month claims run-out from the end of the performance period.

In addition, the MSPB-Hospital measure does account for the transition to the ICD-10 coding system. Because MSPB-Hospital includes all costs billed during the episode and is not specific to what diagnosis is billed on the claim, the impact of the transition to ICD-10 codes is minimal. The only change required is an update to the MSPB-Hospital risk adjustment model, which utilizes the new version of the CMS-HCC methodology from the Medicare Advantage program that does account for ICD-10 codes.

During the data preview for the MSPB-Hospital measure, each hospital receives a Hospital-Specific Report (HSR) that provides information on the hospital's performance on the MSPB-Hospital measure, as well as three supplementary hospital-specific data files (an index admission file, a beneficiary risk score file, and an MSPB-Hospital episode file) related to the hospital's MSPB-Hospital measure. Together, these files provide an overview of how the hospital performed on the MSPB-Hospital measure as well as a summary of how hospitals in the state and in the nation performed. For example, each hospital's files provide the number of eligible admissions, average spending per episode, MSPB-Hospital amount, and MSPB-Hospital measure for the hospital as well as for the state and the nation. Additionally, each hospital's MSPB-Hospital spending is broken into three categories (i.e., 3 days prior to index admission, during-index admission, and 30 days after hospital discharge), and within these categories, spending levels are broken down by claim type. For comparison, the state and national values for these breakdowns are given to hospitals as well. Further, each hospital's average observed spending and average expected spending (based on beneficiary age and health status) breakdowns by Major Diagnostic Category (MDC) are presented in the hospital's HSR alongside analogous values at the state and national levels to allow the hospital to compare its case mix against the state and the nation. In addition to helping hospitals verify their MSPB-Hospital measure scores and identify opportunities to improve efficiency, providing these files allows us to better communicate MSPB-Hospital scores to hospitals and allows hospitals to provide informed feedback to the measure contractor and CMS. During the 30-day preview periods, the measure contractor and CMS received no reports of errors in the measure's calculation.

[1] Comprehensive Error Rate Testing (CERT) Program. "Appendices Medicare Fee-for-Service 2015 Improper Payments Report". Table A6. https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Medicare-FFS-Compliance-Programs/CERT/CERT-Reports-Items/Downloads/AppendicesMedicareFee-for-Service2015ImproperPaymentsReport.pdf

F.3.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, and algorithm)?**

There are no fees, licensing, or other requirements for use of the MSPB-Hospital measure values and MSPB-Hospital measure spending breakdowns made publicly available on Hospital Compare.

F.3.3. **If there are any fees associated with the use of this measure as specified, attach the fee schedule here. (Save file as: F3_3_FeeSchedule)**

## Usability and Use

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

*NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.*

U.1.1. **Current and Planned Use**

| Specific Plan for Use | Current Use (for current use provide URL) |
|---|---|
| Payment Program | Public Reporting<br>https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/hospitalrhqdapu.html<br>Hospital Inpatient Quality Reporting (IQR)<br>Hospital Compare<br>https://www.medicare.gov/hospitalcompare/search.html<br>https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/hospitalrhqdapu.html<br>Hospital Inpatient Quality Reporting (IQR)<br>Hospital Compare<br>https://www.medicare.gov/hospitalcompare/search.html |

U.1.2. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose

- Geographic area and number and percentage of accountable entities and patients included

HOSPITAL IQR PROGRAM:
Program Name: Hospital Inpatient Quality Reporting (IQR) Program
Sponsor: CMS
Purpose: Section 501(b) of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003 established the Hospital IQR program. The Hospital IQR program pays hospitals that successfully report designated quality measures a higher annual update to their payment rates. The program also provides CMS and the public with data to help consumers make more informed decisions about their health care. Some of the measure information gathered through the Hospital IQR program is available to consumers on the Hospital Compare website at: https://www.medicare.gov/hospitalcompare/search.html. CMS provides each eligible hospital a confidential Hospital-Specific Report (HSR) that provides information on its performance on the MSPB-Hospital measure. These reports, along with the accompanying confidential data files, can be used by hospitals to validate the calculation of their MSPB-Hospital measure values.
Geographic Area: U.S.
Number/Percentage of Accountable Entities: In the FY2017 Hospital IQR Program, which used the MSPB-Hospital measure calculated based on January 1, 2015- December 31, 2015 performance period, 3,207 IQR-eligible hospitals received an MSPB-Hospital measure out of 3,213 IQR-eligible hospitals (99.81%). Additionally, 3,228 hospitals out of 3,298 hospitals eligible to receive an MSPB-Hospital measure score (97.9%) received HSRs for the January 1, 2015 to December 31, 2015 period of performance
Number/Percentage of Patients: N/A
HOSPITAL COMPARE:
Program Name: Hospital Compare
Sponsor: CMS
Purpose: Hospital Compare has information about the quality of care at over 4,000 Medicare-certified hospitals across the country. The public can use Hospital Compare to find hospitals and compare the quality of their care. Specifically, hospitals' MSPB-Hospital measure values will be publicly reported on the Hospital Compare website. However, only hospitals with 25 or more eligible episodes will have their MSPB-Hospital values posted. This requirement reduces the likelihood that a hospital's MSPB-Hospital measure is skewed by a few high- or low-cost episodes. CMS provides each eligible hospital a confidential Hospital-Specific Report (HSR) that provides information on its performance on the MSPB-Hospital measure. These reports, along with the accompanying confidential data files, can be used by hospitals to validate the calculation of their MSPB-Hospital measure values.
Geographic Area: U.S.
Number/Percentage of Accountable Entities and Patients: Please see above as the Hospital Compare public reporting is part of the Hospital IQR Program.

HOSPITAL VBP PROGRAM:
Program Name: Hospital Value-Based Purchasing (VBP) Program
Sponsor: CMS
Purpose: Section 3001 of the Patient Protection and Affordable Care Act (ACA) established the Hospital Value-Based Purchasing (VBP) program. The Hospital VBP program provides financial incentives to subsection (d) hospitals based on their performance on selected quality measures. Section 1886(o)(2)(B)(ii) of the Social Security Act, 3001 of the Patient Protection and Affordable Care Act requires that CMS implement a measure of Medicare spending per beneficiary as part of it Hospital Value-Based Purchasing (VBP) initiatives. The hospital performance score for a performance period will be determined using a higher of its achievement or improvement score for the MSPB-Hospital measure as described in the FY 2012 IPPS Final Rule at 76 FR 51654-56. The MSPB-Hospital measure score will be incorporated into the Hospital VBP Program as part of the Efficiency domain. Because the MSPB-Hospital measure is the only measure currently in the Efficiency domain, the total points earned for the domain would be the points earned on the MSPB-Hospital measure. Each hospital´s Total Performance Score (TPS), used to calculate each hospital´s incentive payment, is calculated by combining its component domain scores. A hospital's improvement score is calculated from a comparison of the hospital's MSPB-Hospital measure value during a period of performance against the MSPB-Hospital measure value during a baseline period.
Geographic Area: U.S.
Number/Percentage of Accountable Entities: In the FY2016 Hospital VBP Program, 3,036 received the MSPB-Hospital measure out of 3,041 hospitals (99.8%). Note that the number of hospitals represents those that had an MSPB-Hospital measure during the baseline and performance period. The FY2016 Hospital VBP program baseline period for MSPB-Hospital was January 1, 2012 to December 31, 2012 and the performance period was January 1, 2014 to December 31, 2014.
Number/Percentage of Patients: N/A

U.1.3. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program,**

**certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A. This is not applicable as the MSPB-Hospital measure is reported in hospital-specific reports and is included as part of the CMS Hospital Inpatient Quality Reporting (IQR) Program and the Hospital Value-Based Purchasing (VBP) Program.

U.1.4. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A. This is not applicable as the MSPB-Hospital measure is reported in hospital-specific reports and is included as part of the CMS Hospital IQR Program and the Hospital VBP Program.

U.2.1. **Progress on Improvement. (Not required for initial endorsement unless available.) Performance results on this measure (current and over time) should be provided in IM.2.2 and IM.2.4.**
**Discuss:**

- **Purpose Progress (trends in performance results)**
- **Geographic area and number and percentage of accountable entities and patients included**

When comparing MSPB-Hospital measure scores between 2014 and 2015 at the provider level, we see that nearly half of all hospitals improved on their MSPB-Hospital measure score, as discussed in section IM.2.2. The MSPB-Hospital measure is able to effectively capture provider risk-adjusted spending during an episode and is able to capture differences between providers. Results from our testing are described in depth in the Testing Attachment included in this submission. Furthermore, our comparison of provider performance between 2014 and 2015 suggests that providers are reducing their average risk-adjusted episode spending and are improving on their MSPB-Hospital measure score.

U.2.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

U.3.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No unintended consequences to individuals or populations have been identified during testing, and no evidence of unintended negative consequences to individuals or populations have been reported since implementation.

## Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

H.1. **Relation to Other NQF-endorsed Measures**
If there are related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

H.1.1. **List of related or competing measures (selected from NQF-endorsed measures)**

H.1.2. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

**H.2. Harmonization**

H.2.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

**Are the measure specifications completely harmonized?**

H.2.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

H.3. **Competing Measure(s)**

H.3.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
**Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)**
The MSPB-Hospital measure evaluates hospitals' efficiency relative to the efficiency of the median hospital. The target population is Medicare beneficiaries enrolled in Medicare Parts A and B who were discharged from short-term acute hospitals. There are currently no NQF-endorsed measures that address both this same measure focus AND this same target population.

## Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services
Co.2 **Point of Contact:** Kimberly, Spalding Bush, kimberly.spaldingbush@cms.hhs.gov, 410-786-3232-
Co.3 **Measure Developer if different from Measure Steward:** Acumen, LLC
Co.4 **Point of Contact:** Rachel, Liu, hvbp-support@acumenllc.com, 650-558-8882-416

## Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
List the workgroup/panel members' names and organizations.
Describe the members' role in measure development.

**Measure Developer/Steward Updates and Ongoing Maintenance**
Ad.2 **Year the measure was first released:** 2012
Ad.3 **Month and Year of most recent revision:** 06, 2016
Ad.4 **What is your frequency for review/update of this measure?** Yearly
Ad.5 **When is the next scheduled review/update for this measure?** 06, 2017

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**