**National Quality Forum—Composite Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** Click here to enter NQF number

**Composite Measure Title**: Global Malnutrition Composite Measure

**Date of Submission**: 3/6/2020

**Composite Construction:**

Two or more individual performance measure scores combined into one score

All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)

|  |
| --- |
| **Instructions: Please contact NQF staff before you begin.**   * If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure submission. * 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. * **Sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.** * **For composites with outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) and composites (2c) 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 25 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](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. and the 2017 Measure Evaluation Criteria and Guidance. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing 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**](#Note10) 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 **instrument-based measures** (including **PRO-PMs**) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **Validity testing** [**11**](#Note11) 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 **instrument based measures (including** **PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** Exclusions are supported by the clinical evidenceand are of sufficient frequency to warrant inclusion in the specifications of the measure; [**12**](#Note12)  **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**](#Note13)  **2b3.** **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 social risk factors) that influence the measured outcome and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b4.** 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**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** 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.  **2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:**  **2c1.** the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and  **2c2**.the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.  (*if not conducted or results not adequate, justification must be submitted and accepted)*  **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. The degree of consensus and any areas of disagreement must be provided/discussed.  **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. |

**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 different components in the composite, indicate the component after the checkbox. If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.17*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| claims | claims |
| registry | 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*).

**1.3. What are the dates of the data used in testing**? March – October 2018 for Validity Testing; January 2019 – December 2019 for Reliability Testing;

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

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.20*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

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

***Table 1 – Description of Measured Entities Included in Measure Testing***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **State** | **Hospital Type** | **Bed Size** | **Urban/Rural** | **Total Patients Included** |
| 1 | NC | Short Term Acute Care | Medium | Urban | 413 |
| 2 | NC | Short Term Acute Care | Large | Urban | 2064 |
| 3 | NC | Community Hospital | Medium | Urban | 1042 |
| 4 | NC | Academic Medical Center | Large | Urban | 2256 |
| 5 | NC | Academic Medical Center | Large | Urban | 1619 |
| 6 | VA | Critical Access Hospital | Small | Rural | 135 |
| 7 | VA | Critical Access Hospital | Small | Rural | 287 |
| 8 | VA | Short Term Acute Care | Medium | Rural | 281 |
| 9 | VA | Short Term Acute Care | Large | Urban | 3310 |
| 10 | WV | Academic Medical Center | Large | Urban | 1403 |
| 11 | NY | Academic Medical Center | Large | Urban | 1935 |
| 12 | NC | Short Term Acute Care | Large | Urban | 4297 |
| 13 | CO | Academic Medical Center | Large | Urban | 1759 |
| 14 | PA | Short Term Acute Care | Medium | Urban | 1484 |
| 15 | PA | Short Term Acute Care | Small | Rural | 152 |
| 16 | PA | Short Term Acute Care | Medium | Urban | 687 |
| 17 | PA | Short Term Acute Care | Large | Urban | 1363 |
| 18 | PA | Short Term Acute Care | Medium | Urban | 410 |
| 19 | PA | Short Term Acute Care | Medium | Rural | 504 |
| 20 | PA | Short Term Acute Care | Medium | Urban | 362 |
| 21 | PA | Short Term Acute Care | Medium | Urban | 446 |
| 22 | PA | Short Term Acute Care | Large | Urban | 1040 |
| 23 | PA | Short Term Acute Care | Medium | Rural | 415 |
| 24 | PA | Short Term Acute Care | Large | Urban | 1568 |
| 25 | PA | Academic Medical Center | Large | Urban | 1850 |
| 26 | PA | Academic Medical Center | Large | Urban | 2063 |
| 27 | PA | Short Term Acute Care | Medium | Urban | 920 |

**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*).

A total of 37,450 patients age 65 years and older were included in the testing population across 27 acute care hospitals in 6 states. Out of the total, 53.3% were female. In terms of age breakdown, 46% were 65-74, 33.9% were 75-84, and 20.1% were 85+. Race distribution was as follows: 81.8% White, 12% Black, 1.2% Asian or Pacific Islander, 0.2% American Indian or Alaska Native, 1.9% Other, and 2.9% were unable to be determined. Out of the total, 2.2% were of Hispanic ethnicity.

Data quality was a concern for the time-to-screening data point for patients above the 99th percentile and were therefore excluded from the analysis (N=473). The capture of screening data longer than 48 hours prior to admission was not included in the dataset. These screening results are not considered to be clinically reliable according to clinical guidance by expert consensus and as outlined by the Academy of Nutrition and Dietetics. Nutritional status as identified via malnutrition screening should occur during the admission process.

**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**.

A separate and more recent dataset was constructed to complete additional testing for the composite measure reliability. A total of 179,336 patients age 65 years and older were included in the testing population across 56 acute care hospitals in 10 states. This newer dataset was similar in demographic breakdown of the validity testing dataset. Age-wise, the average age was 76.5 and the mean age was 75. In terms of race, the cohort was 77.8% White, 9.68% Black, 1.59% Asian or Pacific Islander, and 9.56% Other. The sample also included 4.91% who were identified as Hispanic.

***Table 1 – Description of Measured Entities Included in Measure Reliability Testing***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **State** | **Hospital Type** | **Bed Size** | **Urban/Rural** | **Total Patients Included** |
| 1 | NC | Short Term Acute Care | Medium | Urban | 27087 |
| 2 | NC | Short Term Acute Care | Large | Urban | 9222 |
| 3 | VA | Critical Access Hospital | Small | Rural | 135 |
| 4 | PA | Short Term Acute Care | Medium | Urban | 3037 |
| 5 | PA | Short Term Acute Care | Small | Rural | 229 |
| 6 | PA | Short Term Acute Care | Medium | Urban | 1556 |
| 7 | PA | Short Term Acute Care | Large | Urban | 2954 |
| 8 | PA | Short Term Acute Care | Medium | Urban | 657 |
| 9 | PA | Short Term Acute Care | Medium | Rural | 965 |
| 10 | PA | Short Term Acute Care | Medium | Urban | 820 |
| 11 | PA | Short Term Acute Care | Medium | Urban | 857 |
| 12 | PA | Short Term Acute Care | Large | Urban | 1887 |
| 13 | PA | Short Term Acute Care | Medium | Rural | 903 |
| 14 | PA | Short Term Acute Care | Large | Urban | 3157 |
| 15 | PA | Academic Medical Center | Large | Urban | 3629 |
| 16 | PA | Academic Medical Center | Large | Urban | 3782 |
| 17 | PA | Short Term Acute Care | Medium | Urban | 1984 |
| 18 | UT | Short Term Acute Care | Small | Urban | 1030 |
| 19 | UT | Short Term Acute Care | Small | Urban | 1294 |
| 20 | ID | Short Term Acute Care | Small | Rural | 518 |
| 21 | UT | Short Term Acute Care | Small | Rural | 800 |
| 22 | UT | Short Term Acute Care | Medium | Urban | 8855 |
| 23 | UT | Critical Access Hospital | Small | Rural | 233 |
| 24 | UT | Short Term Acute Care | Large | Urban | 10785 |
| 25 | UT | Short Term Acute Care | Medium | Urban | 2152 |
| 26 | UT | Short Term Acute Care | Small | Urban | 180 |
| 27 | UT | Short Term Acute Care | Medium | Urban | 1682 |
| 28 | UT | Short Term Acute Care | Large | Urban | 5763 |
| 29 | UT | Short Term Acute Care | Small | Rural | 711 |
| 30 | UT | Short Term Acute Care | Small | Urban | 1000 |
| 31 | UT | Critical Access Hospital | Small | Rural | 175 |
| 32 | UT | Short Term Acute Care | Small | Rural | 386 |
| 33 | UT | Short Term Acute Care | Large | Urban | 6386 |
| 34 | WI | Academic Medical Center | Large | Urban | 6081 |
| 35 | NJ | Academic Medical Center | Large | Urban | 11760 |
| 36 | IA | Short Term Acute Care | Large | Urban | 1560 |
| 37 | IA | Short Term Acute Care | Medium | Urban | 4740 |
| 38 | IA | Short Term Acute Care | Small | Urban | 785 |
| 39 | PA | Short Term Acute Care | Medium | Urban | 1071 |
| 40 | PA | Short Term Acute Care | Medium | Urban | 566 |
| 41 | PA | Short Term Acute Care | Small | Rural | 608 |
| 42 | PA | Short Term Acute Care | Large | Rural | 1551 |
| 43 | PA | Short Term Acute Care | Medium | Urban | 326 |
| 44 | PA | Short Term Acute Care | Small | N/A | 859 |
| 45 | ME | Short Term Acute Care | Medium | Urban | 938 |
| 46 | TX | Short Term Acute Care | Medium | Urban | 1391 |
| 47 | TX | Short Term Acute Care | Small | Urban | 873 |
| 48 | TX | Short Term Acute Care | Medium | Urban | 3845 |
| 49 | TX | Short Term Acute Care | Medium | Urban | 3203 |
| 50 | TX | Short Term Acute Care | Medium | Urban | 6280 |
| 51 | TX | Short Term Acute Care | Medium | Urban | 3824 |
| 52 | TX | Short Term Acute Care | Medium | Urban | 5449 |
| 53 | TX | Short Term Acute Care | Medium | Urban | 2036 |
| 54 | TX | Short Term Acute Care | Large | Urban | 4506 |
| 55 | TX | Short Term Acute Care | Large | Urban | 5348 |
| 56 | TX | Short Term Acute Care | Large | Urban | 7061 |

**1.8 What were the social risk factors that were available and analyzed?** For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

No social risk factor data were collected for testing.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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*)

***Note****: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.*  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. 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*)

Composite measure reliability was assessed using the variance components—extracted from a linear mixed effects (LME) model—to calculate the intraclass correlation coefficient (ICC). The LME framework was employed, because it accommodates inclusion of both fixed and random effects, the latter of which account, statistically, for the correlated or non-independent nature of measures that are hierarchically nested within health systems (N = 10) and practice sites (N = 56). The model variance (σ2) can then be partitioned into components that are, in turn, used to calculate the ICC.

Drawing on this well-established framework, a three-step process was followed to calculate the ICC. First, an intercept-only LME model was fitted to the composite measure data, incorporating health system (HSYSTEM) as a random intercept term. Second, the between-system (σ2between) and within-system (σ2within) were extracted from the LME model, and, third, the ICC was generated by taking the ratio of the respective variance components:

ICC = σ2between/(σ2between + σ2within).

The reliability assessment was carried using the **lme4** and **performance** packages in R. Specifically, the **lme4** package was used to fit of the LME model to the composite measure data, and the **performance** package—a bundle of utility functions for assessing statistical model quality—was used to capture the model variance components and ICC.

**2a2.3. 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 ICC was calculated in two models, without case minimums (Model 1) and with the following exclusion criteria as reflected in the measure specifications (Model 2):

1. A minimum of 20 cases in the denominator for each measures
2. A minimum of three reportable measures given case minimum as described in 1.

**Model 1 (Without Case Minimums)**

*Number of Observations: 56, Groups (HSYSTEM): 10*

*Syntax*

Linear mixed model fit by REML ['lmerMod']

Formula: MEAS ~ 1 + (1 | HSYSTEM)

Data: data1

*Output*

REML criterion at convergence: 57.5

Scaled residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum | 1Q | Median | 3Q | Maximum |
| -3.5799 | -0.2062 | 0.0798 | 0.4261 | 1.5887 |

Random effects:

|  |  |  |
| --- | --- | --- |
| Groups Name | Variance | Standard Deviation |
| HSYSTEM | 0.2090 | 0.4571 |
| Residual | 0.1139 | 0.3375 |

Fixed effects:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate | Standard Error | t-value |
| (intercept) | 3.1165 | 0.1627 | 19.15 |

*Result*

ICC: 0.647

**Model 2 (With Case Minimums)**

*Number of Observations: 47, Groups (HSYSTEM): 10*

*Syntax*

Linear mixed model fit by REML ['lmerMod']

Formula: MEAS ~ 1 + (1 | HSYSTEM)

Data: data2

*Output*

REML criterion at convergence: 6

Scaled residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum | 1Q | Median | 3Q | Maximum |
| -3.3660 | -0.3503 | 0.0652 | 0.4144 | 2.6385 |

Random effects:

|  |  |  |
| --- | --- | --- |
| Groups Name | Variance | Standard Deviation |
| HSYSTEM | 0.18923 | 0.4350 |
| Residual | 0.03623 | 0.1903 |

Fixed effects:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate | Standard Error | t-value |
| (intercept) | 3.1306 | 0.1446 | 21.65 |

*Result*

ICC: 0.839

**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?*)

With regard to using a calculation of intraclass correlation (ICC) to detect signal to noise, a reliability score of 0.70 or greater is considered acceptable for drawing conclusions about groups. The measure’s reliability was tested with and without cases minimums typically recommended by CMS in its quality reporting programs in order to demonstrate the measure’s reliability with those case minimums in place. With case minimums, the ICC calculated was 0.839 and without case minimums it resulted in an ICC of 0.647. This statistic indicates that the composite measure is well within the range established as acceptable for reliability, meaning the composite performance measure score is able to detect meaningful differences among provider groups.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b1. VALIDITY TESTING**

***Note****: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.*

**2b1.1. What level of validity testing was conducted**?

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

**Composite performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**Validity testing for component measures** *(check all that apply)*

***Note****: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

**Endorsed (or submitted) as individual performance measures**

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

**Empirical validity testing of the component measure score(s)**

**Systematic assessment of face validity of component measure score(s) 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*)

**2b1.2. 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)*

*Composite Performance Measure Score Validity Testing*

To empirically test the construct validity of the overall composite measure at the score level, a hierarchical linear regression was conducted to demonstrate that the predictability of the model significantly improved when the components in aggregate were included into the model over standard predictors of these outcomes such as patient characteristics, primary diagnoses, and comorbidities.

The impact of the composite measure components on 30-day readmissions and LOS was assessed using hierarchical regression analysis. Independent variables fell into two categories: “demographic and clinical” and “malnutrition.” A stepwise approach was taken to measure the explanatory power of the malnutrition variables. The hospital 30-day readmissions and LOS models were initially estimated using only the demographic and clinical variables. Next, the models were re-estimated including the malnutrition variables. This approach allowed us to estimate the incremental improvement in goodness-of-fit from including the malnutrition variables. Model goodness-of-fit was reported as adjusted-R2 for the hospital LOS model and the concordance statistic (c-statistic) for the 30-day readmissions model. The statistical significance of the improvement of model fit was tested using the change in -2 residual log-likelihood.

A secondary analysis was conducted to specifically assess the association between the main clinical endpoint of the composite measure (nutrition care plans for patients with a diagnosis of malnutrition) and the outcomes most associated with malnutrition (30-day readmissions and length of stay). The analysis intended to understand the association of having a nutrition care plan with a malnutrition diagnosis vs not having a nutrition care plan.

*Validity Testing for Component Measures – Critical Data Elements*

Construct validity of the critical data elements for the component measures was tested by developing a generalized linear (logistic) regression model. The response variable was Medical Diagnosis (2 levels) as it is the logical outcome of proper screening and assessment for malnutrition. Predictor variables were Screening Result (3 levels), Time to Assessment (3 levels) and Assessment Result (3 levels). An additional test was conducted to ensure the overall linear model for predicting diagnosis was also predictive of the nutrition care plan. The hypothesis for this test is that all predictor variables would be correlated to the outcome of malnutrition diagnosis and that together they would be a strong predictor of the malnutrition outcome, supporting the validity of including these components in the malnutrition composite.

In addition to testing the components of the measure for validity towards the outcome of the composite measure, testing was completed to assess the correlation between the components and outcome of the composite measure with clinical outcomes of patient length of stay (LOS) and 30-day readmissions. This phase of testing assessed the predictive relationship between the set of measure components and LOS and Readmissions, adjusting for differences in patient characteristics. A generalized linear mixed model approach was utilized to conduct the analyses, also known as hierarchical linear modeling.

In the description of the models, the following notation is used:

***Table 2 – Variables Included in the Model***

|  |  |
| --- | --- |
|  | Response (dependent) variable for Patient *i* treated in Site *j*   1. LOS 2. Readmission (yes/no) |
|  | Overall intercept |
|  | Main effects for *k* explanatory (predictor) variables   * Patient-level predictors (patient sex, race, and Hispanic ethnicity)  1. Screening result (2 categories: At-risk, Not-at-Risk) 2. Time-to-Assessment (3 categories: median split & none) 3. Medical Diagnosis of Malnutrition (2 categories: Yes/No) 4. Nutrition Care Plan (2 categories: Yes/No) 5. Primary Diagnosis (CCS-2 Category Level) |
|  | Two-way interactions of predictor variables |
|  | Site-specific random effect |
|  | Individual error term |

The model can be defined as:

The random effect parameter, *µj*, is included to account for the non-independence of data from patients treated in the same facility. This controls for the different (and unmeasured) characteristics of the separate treatment sites.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

*Composite Performance Measure Score Validity Testing*

Our major finding is that malnutrition indicators are significantly related to LOS and Readmissions after controlling for the other variables that were included in the model (patient demographics and primary diagnosis) known to be predictive of those outcomes. The R2 statistic for the LOS model was 0.063 prior to the inclusion of the aggregate measure components and 0.288 after (p<0.001), and the c-statistic for the 30-day readmissions model was 0.614 before their inclusion and 0.625 after (p<0.01).

However, to better characterize the predictability of our current malnutrition outcomes model for length of stay and readmissions, we sought to compare the predictability of CMS’ HCC risk-adjustment model. The HCC model predicts total annual costs, and the statistical models which were evaluated by RTI in 2011 demonstrated the predictive ability for individuals of prospective diagnosis-based models had R2 values ranging from 0.0186 to 0.1246 (evaluated by RTI in 2011). Given the statistics shared above, the strength of predictability of this model and overall measure is adequate and comparable to those already being implemented by CMS for similar purposes.

The secondary analysis of the relationship between a documented nutrition care plan and risk of 30-day readmissions in patients with a malnutrition diagnosis showed a statistically significant relative risk reduction of 24% (21.4% vs. 26.5%, respectively) in the likelihood of 30-day readmissions (OR=0.74, 99%, CI=0.558-0.941). For LOS, hospitalized patients with a malnutrition diagnosis who had a nutrition care plan had on average, a 3-day longer LOS than malnourished patients without a nutrition care plan (LOS of 9.46 vs. 6.46 days, respectively; *p*=0.0001).

*Validity Testing for Component Measures – Critical Data Elements*

***Table 3 - Results of Generalized Linear Regression Model on Composite Outcome***

|  |  |  |  |
| --- | --- | --- | --- |
| Effect | df | Wald Chi-Square | p-value |
| Screening Result | 2 | 75.1 | <.0001 |
| Time to Assessment | 2 | 1094.5 | <.0001 |
| Assessment Result | 2 | 2006.8 | <.0001 |
| Screening Result \* Time to Assessment | 4 | 480.9 | <.0001 |
| Screening Result \* Assessment Result | 4 | 609.0 | <.0001 |
| **Effect** | **df** | **Fisher’s Exact Test** | **p-value** |
| Malnutrition Diagnosis \* Nutrition Care Plan | 1 | 7584.5 | < .0001 |
| c-statistic: 0.828 [fit of the overall score] | | | |

***Table 4 - Length of Stay Predictability of Malnutrition Composite Measure Components***

\*TTA = Time to Assessment; The timing was tested at the median split for all hospitals included in the testing dataset.

***Table 5 – 30-Day Readmissions Predictability of Malnutrition Composite Measure Components***

**2b1.4. 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?*)

*Composite Performance Measure Score Validity Testing*

As reported in the results of both analyses, the composite measure results are strongly correlated to important clinical outcomes associated with malnutrition in the literature, 30-day readmissions and length of stay. Furthermore, the secondary analysis demonstrated that nutrition care plans may be associated with a reduced risk of 30-day readmission for those with malnutrition vs those who are diagnosed with malnutrition but do not have a nutrition care plan.

*Validity Testing for Component Measures – Critical Data Elements*

As outlined in Table 3, all main effects and 2 2-way interactions were highly significant (all p-values <.0001), consistent with our hypotheses. The c-statistic of 0.828 indicates an excellent fit of the model to the malnutrition diagnosis and nutrition care plan. A c-statistic above 0.8 normally indicates a very strong predictive model.

The results in Tables 4 and 5, demonstrate that all components of the malnutrition composite measure, including the outcome of the malnutrition composite measure (malnutrition diagnosis and nutrition care plan) were significantly predictive of the outcome of length of stay (p<0.0001) and 30-day readmissions (p<0.0001).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. EXCLUSIONS ANALYSIS**

***Note:*** *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b2.1. 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 two main exclusions for this measure are a length of stay <24 hours as those patients are not in the hospital long enough to receive proper care for malnutrition. Patients who are transferred or discharged to hospice have significantly different requirements for nutrition support and those treatment plans are highly dependent on patient preferences.

Our project team tested measure exclusion criteria for both their impact on the measure performance score and validity statistics for each individual component measure when they were first developed. The project team tested the measure specifications with a set of hypothetical measure exclusions that were determined by consensus agreement of the Technical Expert Panel but were not explicitly identified in the evidence review. We assessed the measure performance score of each respective testing site with the exclusion criteria and without in order to determine the exclusion criteria’s impact on the facility’s score.

**2b2.2. 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*)

In the original measure testing of the individual components, we identified that neither of the exclusion criteria had significant impacts on the performance scores. When measures were constructed with and without exclusions no p-values reached significance when a two-tailed t-test was performed on the difference between the performance scores.

|  |  |
| --- | --- |
| Component Measure | t-test p-value |
| Malnutrition Screening | p>0.3 |
| Nutrition Assessment | p>0.4 |
| Malnutrition Diagnosis | p>0.8 |
| Nutrition Care Plan | p>0.3 |

**2b2.3. 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*)

At the individual measure level, there was no significant impact of the measure exclusions on the performance measure scores for all 4 component measures.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***Note:***  *Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.*  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b4***](#section2b5)***.***

**2b3.1. What method of controlling for differences in case mix is used?** *(check all that apply)*

**Endorsed (or submitted) as individual performance measures**

**No risk adjustment or stratification**

**Statistical risk model with** Click here to enter number of factors **risk factors**

**Stratification by** Click here to enter number of categories **risk categories**

**Other,** Click here to enter description

**2b3.1.1 If using statistical risk models, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

**2b3.2. 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**.

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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*) **Also discuss any “ordering” of risk factor inclusion;** for example, are social risk factors added after all clinical factors?

**2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:**

**Published literature**

**Internal data analysis**

**Other (please describe)**

**2b3.4a. What were the statistical results of the analyses used to select risk factors?**

**2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk 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.) Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.

**2b3.5. 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*)

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b3.9***](#question2b49)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

**2b3.9. Results of Risk Stratification Analysis**:

**2b3.10. 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*)

**2b3.11.** **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*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

***Note:*** *Applies to the composite performance measure.*

**2b4.1****.** **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)*

A bootstrap resampling methodology was employed to generate a 95% confidence interval around the composite score mean. The 95% confidence interval will then be used to group providers into performance categories (Low, Medium, High).

Specifically, the **resample** package in R was used to generate a bootstrap sample (N = 10,000) from the empirical, i.e. observed, distribution of composite scores across providers. The mean score and standard error—derived from the bootstrap distribution—were estimated, and a 95% CI was generated to drive categorization of provider performance above, within and below the 95% CI of the mean score estimate.

Participating hospitals were categorized into three tiers that reflect those whose composite measure performance scores were above, overlapped with, or were below the 95% estimate generated in the bootstrap analysis. If a hospital’s composite score was assigned a Tier 3 score it was above the estimated confidence interval and implies that the specific hospital’s performance was above the average of the estimate developed from the aggregate of all reporting sites. A hospital receiving a Tier 2 score means their performance was not meaningfully different than the estimated mean. Finally, a hospital receiving a Tier 1 score implies that their composite performance score fell below the mean estimate interval reflective of lower than expected performance.

**2b4.2. 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*)

Among hospitals that meet the case minimum of 20 patients and at least 3 reportable measures, 44.7% of hospitals were in the highest performing Tier 3, 14.9% were in Tier 2, and 40.4% were in Tier 1.

**January 1, 2019 through December 31, 2019**

|  |  |  |
| --- | --- | --- |
|  | **All Participants** | **Participants N≥ 20** |
| **Category** | **Number of Hospitals** | **Number of Hospitals** |
| **Tier 3** | 22, 39.3% | 21, 44.7% |
| **Tier 2** | 3, 5.4% | 7, 14.9% |
| **Tier 1** | 31, 55.3% | 19, 40.4% |

**2b4.3. 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?*)

This tiering approach informed by the bootstrap sample derived from the observed performance measures was used to appropriately distinguish sites with varying degrees of performance among the component measures. These differences ultimately translated to variation in performance on the overall composite measure. Our specific sample of sites is relatively homogeneous because the participating hospitals have been targeting improvement on these quality measures for 1-3 years.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***Note:***  *Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.*

***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 social risk 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 social risk 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.***

**2b5.1. 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*)

**2b5.2. 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*)

**2b5.3. 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?*)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

***Note:***  *Applies to the overall composite measure.*

**2b6.1. 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*)

In the database integration process, a test of consistency for the core data elements is conducted to ensure that the main care processes measured (malnutrition screening and nutrition assessment) are not missing data. For instance, if an assessment is performed, data on the assessment result and the time interval between the screening and assessment copmletion is also present. A consistency measure reflective of the presence of these data corresponding to measured care processes was calculated to see the rate of observed care processes vs expected number of results and time intervals as appropriate.

**2b6.2. 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*)

The average consistency measure across the sample of hospitals in the testing dataset was >95%.

**2b6.3. 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*)

The rates of missing data were consistently low across all reporting sites due to very high feasibility of the data elements as these data are collected during the care process and do not introduce any burden to clinicians. Due to these factors and the consistency statistic results, we conclude that systematic missing data is not biasing performance for this measure.

**2c. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH**

***Note:*** *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

**2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.**

**2d1.1 Describe the method used** (*describe the steps―do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

See component-level measure validity testing described in section 2b1. In this section, we demonstrate how each component contributes to both the quality construct of the composite measure as well as to two patient indicators that are well-measured and correlated with malnutrition.

However, the first-line analyses for length of stay and 30-day readmissions were conducted to compare patients meeting numerator characteristics of the component measures included in the composite measure. Once the predictability of the outcomes of interest was established additional testing was completed to determine difference in outcomes between patients who were diagnosed and had a care plan versus those that did not (i.e., met versus failed the composite measure). The hypothesis for this cross-tabulation analysis was that the outcome of the composite measure, or malnourished patients with a nutrition care plan, should have a reduced likelihood of readmission than those without a care plan. It was also assumed that length of stay would be higher in patients at-risk of malnutrition and/or diagnosed with malnutrition because patients who are at-risk or malnourished tend to have higher acuity than the general population.

**2d1.2. What were the statistical results obtained from the analysis of the components?** (e.g., *correlations, contribution of each component to the composite score, etc*.; *if no empirical analysis, identify the components that were considered and the pros and cons of each*)

See 2b1.2 for component-level measure validity testing.

Patients (65+) with a malnutrition diagnosis and nutrition care plan had a 24% lower likelihood of 30-day hospital readmissions (21.4% vs. 26.5%, respectively) compared to those without a care plan (OR=0.74, 99%, CI=0.558-0.941).

**2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite?** (i*.e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)*

The results of the validity testing at the both the component and overall composite level support the inclusion of each of the component measures into the composite measure. Each component is indepdently associated with the quality construct and is predictive of outcomes of interest. In aggregate, the components together are better predictors of important patient outcomes of care than just patient characteristics alone.

The results of this cross-tabulation support the association of the malnutrition composite score outcome (malnutrition diagnosis and nutrition care plan) with an important clinical outcome of 30-day readmissions.

**2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible**

**2d2.1 Describe the method used** (*describe the steps―do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

Tests of internal consistency (Chronbach’s alpha and item-to-total correlatiosn) were completed to confirm the equal weighting of each of the measure component’s contribution to the total composite score.

**2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules?** (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

|  |  |  |
| --- | --- | --- |
| Measure | Correlation with Total | Chronbach’s Alpha |
| Component Measure #1 | 0.49 | 0.77 |
| Component Measure #2 | 0.68 | 0.67 |
| Component Measure #3 | 0.59 | 0.72 |
| Component Measure #4 | 0.58 | 0.73 |

**2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct?** (i*.e., what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting*)

Given the acceptable item-to-total correlations and strong internal consistency indicative of how closely related the components are to the total score, we concluded that no differences in weighting are necessary for each component measure at this time.