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

**Measure Number** (*if previously endorsed*)**:** 1550

**Measure Title**: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

**Date of Submission**: 11/3/2020

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process *(including Appropriate Use)* | 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, 2b1, 2b2, and 2b4 must be completed.** * **For 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) 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. |

|  |
| --- |
| **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 evidence and 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.  **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 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: Medicare Enrollment Data (including the Master Beneficiary Summary File) | other: Census Data/American Community Survey, Medicare Enrollment Data (including the Master Beneficiary Summary File) |

**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*).

The datasets used for testing included Medicare Parts A and B claims, as well as the Medicare Enrollment Database (EDB). Additionally, the American Community Survey census data as well as enrollment data were used to assess socioeconomic factors (dual eligible variable obtained through enrollment data; Agency for Healthcare Research and Quality [AHRQ] socioeconomic status [SES] index obtained through census data). The dataset used varies by testing type; see Section 1.7 for details.

**1.3. What are the dates of the data used in testing**? The dates used vary by testing type; see Section 1.7 for details.

**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*)

For this measure, hospitals are the measured entities. All non-federal, short-term acute care inpatient US hospitals (including territories) with Medicare fee-for-service (FFS) beneficiaries aged 65 years and older are included. The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details.

**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*)

The number of admissions/patients varies by testing type: see Section 1.7 for details.

**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**.

The datasets, dates, number of measured hospitals, and number of admissions used in each type of testing are in Table 1.

Measure Development

For measure development, we used Medicare administrative claims data (2008). The dataset also included administrative data on each patient for the 12 months prior to the index admission and the 90 days following it. The dataset contained inpatient and facility outpatient claims and Medicare enrollment database (EDB) data. We randomly split the data (2008) into two equal samples: **the Development Dataset** and **Internal Validation Dataset.**

**Measure Testing**

For analytical updates for this measure, we used three-years of Medicare administrative claims data (April 2016 – March 2019). The dataset also included administrative data on each patient for the 12 months prior to the index admission and the 90 days following it. The dataset contained inpatient and facility outpatient claims and Medicare enrollment database (EDB) data.

Table 1. Dataset Descriptions

| Dataset | Applicable Section in the Testing Attachment | Description of Dataset |
| --- | --- | --- |
| Development and Validation Datasets (Medicare Fee-For-Service Administrative Claims Data) | Section 2b3 Risk Adjustment/Stratification  **2b3.6. Statistical Risk Model Discrimination Statistics**  **2b3.7. Statistical Risk Model Calibration Statistics** | Entire Cohort:  Dates of Data: 2008  Number of admissions = 290,329  Patient Descriptive Characteristics:  mean age = 75.2 years; % male = 35.7%  Number of measured hospitals: 3,223  This cohort was randomly split for initial model testing.  First half of split sample  -Number of Admissions: 145,206  -Number of Measured Hospitals: 3,221  Second half of split sample  -Number of Admissions: 145,123  -Number of Measured Hospitals: 3,223 |
| Testing Dataset Medicare Fee-For-Service Administrative Claims Data (April 1, 2016 – March 30, 2019) | Section 2a2 Reliability Testing  Section 2b1 Validity Testing  Section 2b2 Testing of Measure Exclusion  Section 2b3 Risk Adjustment/Stratification  **2b3.6. Statistical Risk Model Discrimination Statistics**  Section 2b4 Meaningful Differences | Dates of Data: April 2016 – March 2019  Number of admissions = 962,744  Patient Descriptive Characteristics:  meanage = 73.9 years; % male = 37.2  Number of measured hospitals: 3,418 |
| The American Community Survey (ACS) | Section 2b3: Risk adjustment/Stratification for Outcome or Resource Use Measures | Dates of Data: 2013-2017  We used the AHRQ SES index score derived from the American Community Survey (2013-2017) to study the association between the 90-day complication outcome and SRFs. The AHRQ SES index score is based on beneficiary 9-digit zip code level of residence and incorporates 7 census variables found in the American Community Survey. |
| Master Beneficiary Summary File (MBSF) | Section 2b3: Risk adjustment/Stratification for Outcome or Resource Use Measures | Dates of Data: April 2016 – March 2019  We used dual eligible status (for Medicare and Medicaid) derived from the MBSF to study the association between the 90-day measure outcome and dual-eligible status. |

**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.

We selected social risk factor (SRF) variables to analyze after reviewing the literature and examining available national data sources. We sought to find variables that are consistently captured in a reliable fashion for all patients in this measure. There is a large body of literature linking various SRFs to worse health status and higher complications over a lifetime. Income, education, and occupation are the most commonly examined SRFs studied. The causal pathways for SRF variable selection are described below in Section 2b3.3a. Unfortunately these variables are not available at the patient level for this measure. Therefore proxy measures of income, education level and economic status were selected.

The SRF variables used for analysis were:

* Dual eligible status: Dual eligible status (in other words, being enrolled in both Medicare and Medicaid) patient-level data is obtained from the CMS Master Beneficiary Summary File (MBSF)

Following guidance from ASPE and a body of literature demonstrating differential health care and health outcomes among dual eligible patients, we identified dual eligibility as a key variable (ASPE 2016; ASPE 2020). We recognize that Medicare-Medicaid dual eligibility has limitations as a proxy for patients' income or assets because it does not provide a range of results and is only a dichotomous outcome. However, the threshold for over 65-year-old Medicare patients is valuable, as it takes into account both income and assets and is consistently applied across states for the older population. We acknowledge that it is important to test a wider variety of SRFs including key variables such as education and poverty level; therefore, we also tested a validated composite based on census data linked to as small a geographic unit as possible.

* AHRQ-validated SES index score (summarizing the information from the following 7 variables): percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12th grade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room)

Finally, we selected the AHRQ SES index score because it is a well-validated variable that describes the average SES of people living in defined geographic areas (Bonito et al., 2008). Its value as a proxy for patient-level information is dependent on having the most granular-level data with respect to communities that patients live in. We considered the area deprivation index (ADI) among many other potential indicators when we initially evaluated the impact of SDS indicators. We ultimately did not include the ADI at the time, partly due to the fact that the coefficients used to derive ADI had not been updated for many years. Recently, the coefficients for ADI have been updated and therefore we compared the ADI with the AHRQ SES Index and found them to be highly correlated. In this submission, we present analyses using the census block level, the most granular level possible using American Community Survey (ACS) data. A census block group is a geographical unit used by the US Census Bureau which is between the census tract and the census block. It is the smallest geographical unit for which the bureau publishes sample data. The target size for block groups is 1,500 and they typically have a population of 600 to 3,000 people. We used 2013-2017 ACS data and mapped patients’ 9-digit ZIP codes via vendor software to the census block group level. Given the variation in cost of living across the country, the median income and median property value components of the AHRQ SES Index were adjusted by regional price parity values published by the Bureau of Economic Analysis (BEA). This provides a better marker of low SES neighborhoods in high expense geographic areas. We then calculated an AHRQ SES Index score for census block groups that can be linked to 9-digit ZIP codes. We used the percentage of patients with an AHRQ SES index score equal to or below 42.7 to define the lowest quartile of the AHRQ SES Index.

References:

Adler NE, Newman K. Socioeconomic disparities in health: pathways and policies. Health affairs (Project Hope). 2002; 21(2):60-76.

Bonito A, Bann C, Eicheldinger C, Carpenter L. Creation of new race-ethnicity codes and socioeconomic status (SES) indicators for Medicare beneficiaries. Final Report, Sub-Task. 2008;2.

Courtney M, Huddleston J, Iorio R, Markel D. Socioeconomic Risk Adjustment Models for Reimbursement Are Necessary in Primary Total Joint Arthroplasty. July 2016; 32(1):1-5. <https://doi.org/10.1016/j.arth.2016.06.050>.

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Report to Congress: Social Risk factors and Performance Under Medicare’s Value-based Payment Programs. 2016; <https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs>. Accessed November 10, 2019.

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare’s Value-based Purchasing Programs. 2020; <https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf>. Accessed July 2, 2020.

Martsolf G, Barrett M, Weiss A, Kandrack R, Washington R, Steiner C, Mehrotra A, SooHoo N, Coffey R. Impact of Race/Ethnicity and Socioeconomic Status on Risk-Adjusted Hospital Readmission Rates Following Hip and Knee Arthroplasty, The Journal of Bone and Joint Surgery. 2016;98(16):1385-1391. <https://doi.org/10.2106/JBJS.15.00884>.

White, R.S., Sastow, D.L., Gaber-Baylis, L.K. *et al.* Readmission Rates and Diagnoses Following Total Hip Replacement in Relation to Insurance Payer Status, Race and Ethnicity, and Income Status. *J. Racial and Ethnic Health Disparities* 5, 1202–1214 (2018). <https://doi.org/10.1007/s40615-018-0467-0>.

Xu HF, White RS, Sastow DL, Andreae MH, Gaber-Baylis LK, Turnbull ZA. Medicaid insurance as primary payer predicts increased mortality after total hip replacement in the state inpatient databases of California, Florida and New York. *J Clin Anesth*. 2017;43:24‐32. doi:10.1016/j.jclinane.2017.09.008.

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

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

Measure Score Reliability

We performed two types of reliability testing. First, we estimated the overall measure score reliability by calculating the intra-class correlation coefficient (ICC) using a split sample (in other words, test-retest) method. Second, we estimated the facility-level reliability (signal-to-noise reliability).

Split-Sample 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. Accordingly, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produce 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, and then measured again using a second random subset exclusive of the first, and the agreement of the two resulting performance measures compared across hospitals (Rousson, Gasser, and Seifert, 2002).

For split-sample reliability of the measure in aged 65 years and older, we randomly sampled half of patients within each hospital for a three year period, calculated the measure for each hospital, and repeated the calculation using the second half. Thus, each hospital is measured twice, but each measurement is made using an entirely distinct set of patients. To the extent that the calculated measures of these two subsets agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement we calculated the intra-class correlation coefficient (Shrout & Fleiss, 1979), and assessed the values according to conventional standards (Landis & Koch, 1977). Specifically, we used a combined 2016-2019 sample, randomly split it into two approximately equal subsets of patients, and calculated the RSCR for each hospital for each sample. The agreement of the two RSCRs was quantified for hospitals in each sample using the intra-class correlation as defined by ICC (2,1). (Shrout & Fleiss, 1979)

Using two non-overlapping random samples provides a conservative estimate of the measure’s reliability, compared with using two random but potentially overlapping samples which would exaggerate the agreement. Moreover, because our final measure is derived using hierarchical logistic regression, and a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less 'signal', a split sample using a single measurement period would introduce extra noise. This leads to an underestimate in the actual test-retest reliability that would be achieved if the measure were reported using the full measurement period, as evidenced by the Spearman Brown prophecy formula (Spearman 1910, Brown 1910). We used this formula to estimate the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort.

Signal-to-Noise

We estimated the signal to noise reliability (facility-level reliability), which is the reliability with which individual units (hospitals) are measured. While test re-test reliability is the most relevant metric from the perspective of overall measure reliability, it is also meaningful to consider the separate notion of “unit” reliability, that is, the reliability with which individual units (here, hospitals) are measured. The reliability of any one facility’s measure score will vary depending on the number of patients admitted for an elective THA/TKA procedure. Facilities with more volume (in other words, with more patients) will tend to have more reliable scores, while facilities with less volume will tend to have less reliable scores. Therefore, we used the formula presented by Adams and colleagues (2010) to calculate facility-level reliability.

Where facility-to-facility variance is estimated from the hierarchical logistic regression model, n is equal to each facility’s observed case size, and the facility error variance is estimated using the variance of the logistic distribution (π^2/3). The facility-level reliability testing is limited to facilities with at least 25 admissions for public reporting.

Signal to noise reliability scores can range from 0 to 1. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real difference in performance.

Additional Information

In constructing the measure, we aim to utilize only those data elements from the claims that have both face validity and reliability. We avoid the use of fields that are thought to be coded inconsistently across providers. Specifically, we use fields that are consequential for payment and which are audited. We identify such variables through empiric analyses and our understanding of CMS auditing and billing policies and seek to avoid variables which do not meet this standard.

In addition, 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.

Furthermore, we assessed the variation in the frequency of the variables over time: Detailed information is presented in the measure’s 2020 Condition-Specific Measure Updates and Specifications Report cited below.

References:

Adams J, Mehrota, A, Thoman J, McGlynn, E. (2010). Physician cost profiling – reliability and risk of misclassification. NEJM, 362(11): 1014-1021.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

DeBuhr J, McDowell, K, Grady J, et al., 2020 Condition-Specific Complication Measure Updates and Specifications Report - Available at: <https://www.qualitynet.org/inpatient/measures/complication/methodology>.

Landis J, Koch G, The measurement of observer agreement for categorical data, Biometrics, 1977;33:159-174.

Rousson V, Gasser T, Seifert B. "Assessing intrarater, interrater and test–retest reliability of continuous measurements," Statistics in Medicine, 2002, 21:3431-3446.

Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 1979, 86, 420-3428.

Spearman, Charles, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271–295.

**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*)

Measure Score Reliability Results

Signal-to-Noise

We calculated the signal-to-noise reliability score for each hospital with at least 25 admissions\* (see Table 2 below). The median reliability score was 0.87, ranging from 0.46 to 1.00. The 25th and 75th percentiles were 0.74 and 0.94, respectively. The median reliability score demonstrates moderate reliability.

Table 2. Signal-to-noise reliability distribution for THA/TKA complications

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mean** | **Std. Dev.** | **Min** | **5th Percentile** | **10th Percentile** | **25th Percentile** | **Median** | **75th Percentile** | **90th Percentile** | **95th Percentile** | **Max** |
| 0.83 | 0.14 | 0.46 | 0.53 | 0.60 | 0.74 | 0.87 | 0.94 | 0.97 | 0.98 | 1.00 |

\*Hospital measure scores are calculated for all hospitals (including those that have fewer than 25 procedures) but only publicly reported for those that have at least 25 procedures to ensure hospital results are reliable.

Split-Sample Reliability

In total, 962,744 admissions were included in the analysis, using 3 years of data. After randomly splitting the sample into two halves, there were 480,496 admissions from 3,365 hospitals in one half and 482,248 admissions from 3,418 hospitals in the other half. As a metric of agreement, we calculated the ICC for hospitals with 25 admissions or more. Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSCR for each hospital was 0.524.

**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?*)

Measure Score Reliability Results

Using the approach used by Adams et. al. and Yu et al., we obtained the median signal-to-noise reliability score of 0.87, which demonstrates “almost perfect” agreement.

Our interpretation of the results is based on the standards established by Landis and Koch (1977):

< 0 – Less than chance agreement;

0 – 0.2 Slight agreement;

0.21 – 0.39 Fair agreement;

0.4 – 0.59 Moderate agreement;

0.6 – 0.79 Substantial agreement;

0.8 – 0.99 Almost Perfect agreement; and

1 Perfect agreement

The split-sample reliability score of 0.524, discussed in the previous section, represents the lower bound of estimate of the true measure reliability.

In the absence of empirically supported standards, our position is that ‘acceptability’ depends on context. For simple concepts or constructs, such as a patient’s weight, the expectation is that the test-retest reliability of a measure of that construct should be quite high. However, for complex constructs, such as clinical severity, patient comorbidity, or symptom profiles used to identify a condition or clinical state, reliability of measures used to define these constructs is quite a bit lower.

Taken together, these results indicate that there is substantial reliability in the measure score.

References:

Adams J, Mehrota, A, Thoman J, McGlynn, E. (2010). Physician cost profiling – reliability and risk of misclassification. NEJM, 362(11): 1014-1021.

Landis J, Koch G. The measurement of observer agreement for categorical data, Biometrics 1977;33:159-174.

Yu, H, Mehrota, A, Adams J. (2013). Reliability of utilization measures for primary care physician profiling. Healthcare, 1, 22-29.

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

**2b1. VALIDITY TESTING**

**2b1.1. 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 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.

**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)*

Empirical Validity

Stewards of NQF-endorsed measures going through the re-endorsement process are required to demonstrate external validity testing at the time of maintenance review, or if this is not possible, justify the use of face validity only. To meet this requirement for the THA/TKA complications measure, we identified and assessed the measure’s correlation with other measures that target the same domain of quality (e.g. complications, safety, or post-procedure utilization) for the same or similar populations. The goal was to identify if better performance in this measure was related to better performance on other relevant structural or outcomes measures. After literature review and consultations with measures experts in the field, there were very few measures identified that assess the same domains of quality. Given that challenge, we selected the following to use for validity testing:

1. Overall Hospital Star Rating: CMS’s Overall Hospital Star Rating assesses hospitals’ overall performance (expressed on Hospital Compare graphically, as stars) based on a weighted average of “group scores” from different domains of quality (mortality, readmissions, safety, patient experience, imaging, effectiveness of care, timeliness of care). Each group has within it, measures that are reported on Hospital Compare, including this complications measure. Group scores for each individual group are derived from latent-variable models that identify an underlying quality trait for each group. Group scores are combined into an overall hospital score using fixed weights; overall hospital scores are then clustered, using k-means clustering, into five groups and are assigned one-to-five stars (the hospital’s Star Rating). For the validity testing presented in this testing form, we used hospital’s Star Ratings from 3,418 Medicare FFS hospitals from March 2019. The full methodology for the Overall Hospital Star Rating can be found at: <https://www.qualitynet.org/outpatient/public-reporting/overall-ratings/resources>.

1. Hospital THA/TKA Surgical Volume: There is evidence that surgical complication rates for providers (both surgeons and hospitals) decline with increasing volume (Sibley et al., 2017; Murphy et al., 2019; Courtney et al., 2018). Thus, we assessed validity of the measure by examining the relationship between volume and the measure score for hospitals. To establish validity, we expect scores to be correlated with case volume at the hospital level.

We examined the relationship of performance between the THA/TKA complication measure scores (RSCRs) and each of these external measures of hospital quality. For the external measures, the comparison was against performance within quartiles of the Star Ratings overall category (1-5 Stars), as well as hospital THA/TKA surgical volume. We predicted the THA/TKA complication measure scores would have a small association with the overall hospital star rating scores, with lower RSCRs associated with better Star ratings. With THA/TKA surgical volume, we assume that lower RSCRs will be moderately associated with higher volume hospitals.

References

Courtney M, Frisch N, Bohl D, Della Valle C. Improving Value in Total Hip and Knee Arthroplasty: The Role of High Volume Hospitals. *The Journal of Arthroplasty.* 2018;33(1):1-5. <https://doi.org/10.1016/j.arth.2017.07.040>.

Murphy WS, Cheng T, Lin B, Terry D, Murphy SB. Higher Volume Surgeons Have Lower Medicare Payments, Readmissions, and Mortality After THA. *Clin Orthop Relat Res*. 2019;477(2):334‐341. doi:10.1097/CORR.0000000000000370.

Sibley R, Charumbhumi, V, Hutzler L, Paoli A, Bosco, J. Joint Replacement Volume Positively Correlates With Improved Hospital Performance on Centers for Medicare and Medicaid Services Quality Metrics. *The Journal of Arthroplasty*. 2017;32(5):1409-1413. <https://doi.org/10.1016/j.arth.2016.12.010>.

Empirical Validation of Claims-Based Definition of Complications

During original measure development we validated the administrative claims-based definition of THA/TKA complication (original model specification) against medical record data (Dataset 2). The primary goal of this validation study was to determine the overall agreement between patients identified as having a complication (or no complication) using claims data compared with those who had a complication (or no complication) documented in the medical record. We conducted a secondary analysis of agreement of individual, specific complications to identify opportunities for measure improvement.

A statistician and practicing rheumatologist conducted a detailed analysis of each abstracted patient record and compared the findings to the patient results found in the claims data. If any disagreement between the medical record abstraction and the claims data was found, the disagreement was documented and explored in further detail. In some instances, we requested that the medical record be re-abstracted in order to confirm the disagreement and/or to obtain more clinical information. Our clinical team also reviewed some medical records to further determine the nature of disagreement.

To determine overall measure agreement, we calculated the percentage of patients for whom both the claims and medical record identified at least one complication or neither identified a complication. For each case where there was a disagreement between the medical record and claims-based measure, we verified and characterized each disagreement. We then conducted a detailed review of all disagreements between the specific complications documented (or not documented) in the claims data and the medical records, even if such disagreements did not result in overall measure disagreement. We then calculated the percentage of patients where the exact complication(s) coded in claims was also documented in the medical record and vice versa (referred to throughout as “one-to-one agreement”).

Validity of Other Claims-Based Measures:

Our team has demonstrated for a number of prior measures the validity of claims-based measures for profiling hospitals by comparing either the measure results or individual data elements against medical records. CMS validated six NQF-endorsed measures currently in public reporting (acute myocardial infarction [AMI], heart failure, and pneumonia mortality and readmission measures) with models that used chart-abstracted data for risk adjustment. Specifically, claims model validation was conducted by building comparable models using abstracted medical record data for risk adjustment for heart failure patients (National Heart Failure data) (Krumholz, Wang, et al. 2006; Keenan et al. 2008), AMI patients (Cooperative Cardiovascular Project data) (Krumholz, Wang, et al. 2006), and pneumonia patients (National Pneumonia Project dataset) (Bratzler et al. 2011). When both models were applied to the same patient population, the hospital risk-standardized rates estimated using the claims-based risk-adjustment models had a high level of agreement with the results based on the medical record model, thus supporting the use of the claims-based models for public reporting.

We have also completed two national, multi-site validation efforts for two procedure-based complications measures (elective primary THA/TKA and implantable cardioverter defibrillator). Both projects demonstrated strong agreement between complications coded in claims and abstracted medical record data.

References:

Bratzler DW, Normand SL, Wang Y, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. PLoS One 2011;6(4):e17401.

Keenan PS, Normand SL, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation 2008;1(1):29-37.

Krumholz HM, Brindis RG, Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. January 24, 2006 2006;113(3):456-462.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation 2006;113(13):1683-92.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. Circulation 2006;113:1693-1701.

National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report http://www.qualityforum.org/projects/Patient\_Outcome\_Measures\_Phases1-2.aspx. Accessed August 19, 2010.

Shahian DM, He X, O’Brien S, et al. Development of a Clinical Registry-Based 30-Day Readmission Measure for Coronary Artery Bypass Grafting Surgery. Circulation 2014; DOI: 0.1161/CIRCULATIONAHA.113.007541. Published online before print June 10, 2014

Validity Indicated by Established Measure Development Guidelines:

We developed this measure in consultation with national guidelines for publicly reported outcomes measures, with outside experts, and with the public. The measure is consistent with the technical approach to outcomes measurement set forth in NQF guidance for outcomes measures (National Quality Forum, 2010), CMS Measure Management System (MMS) guidance, and the guidance articulated in the American Heart Association scientific statement, “Standards for Statistical Models Used for Public Reporting of Health Outcomes” (Krumholz, Brindis, et al. 2006).

Validity as Assessed by External Groups:

Throughout measure development, we obtained expert and stakeholder input via three mechanisms: regular discussions with an advisory working group, a national TEP, and a 15-day public comment period in order to increase transparency and to gain broader input into the measure.

We assembled the working group and held regular meetings throughout the development phase. The working group was tailored for development of this measure and consisted of clinicians and other professionals with expertise in biostatistics, measure methodology, and quality improvement. Working group meetings addressed key issues related to measure development, including weighing the pros and cons of and finalizing key decisions (e.g., defining the measure cohort and outcome) to ensure the measure is meaningful, useful, and well-designed. The working group provided a forum for focused expert review and discussion of technical issues during measure development prior to consideration by the broader TEP.

In addition to the working group, and in alignment with the CMS MMS, we convened a TEP to provide input and feedback during measure development from a group of recognized experts in relevant fields. To convene the TEP, we released a public call for nominations and selected individuals to represent a range of perspectives, including physicians, consumers, and purchasers, as well as individuals with experience in quality improvement, performance measurement, and health care disparities. We held three structured TEP conference calls consisting of presentation of key issues, our proposed approach, and relevant data, followed by open discussion among TEP members.

Following completion of the preliminary model, we solicited public comment on the measure through the CMS website. The public comments were then posted publicly for 30 days. The resulting input was taken into consideration during the final stages of measure development and contributed to minor modifications to the measure.

Finally, NQF previously endorsed this measure in 2012, demonstrating additional external groups’ endorsement of the measure’s validity.

Face Validity as Determined by TEP:

One means of confirming the validity of this measure was face validity assessed by our TEP.

List of TEP Members

1. Mark L. Francis, MD

Professor of Medicine and Biomedical Sciences, Chief, Division of Rheumatology, Department of Internal Medicine, Texas Tech University Health Sciences Center

2. Cynthia Jacelon, PhD, RN, CRRN

Associate Professor, School of Nursing, University of Massachusetts; Association of Rehabilitation Nurses

3. Norman Johanson, MD

Chairman, Orthopedic Surgery, Drexel University College of Medicine

4. C. Kent Kwoh, MD

Professor of Medicine, Associate Chief and Director of Clinical Research, Division of Rheumatology and Clinical Immunology University of Pittsburgh

5. Courtland G. Lewis, MD

American Association of Orthopaedic Surgeons

6. Jay Lieberman, MD

Professor and Chairman, Department of Orthopedic Surgery, University of Connecticut Health Center; Director, New England Musculoskeletal Institute

7. Peter Lindenauer, MD, M.Sc.

Hospitalist and Health Services Researcher, Baystate Medical Center; Professor of Medicine, Tufts University

8. Russell Robbins, MD, MBA

Principal, Mercer's Total Health Management

9. Barbara Schaffer

THA Patient

10. Nelson SooHoo, MD, MPH

Professor, University of California at Los Angeles

11. Steven H. Stern, MD

Vice President, Cardiology & Orthopedics/ Neuroscience, United Healthcare

12. Richard E. White, Jr., MD

American Association of Hip and Knee Surgeons

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

Comparison to Star-Rating Summary Scores

Figure 1 shows the Box-whisker plots of the THA/TKA complications measure RSCRs within each quartile of Star-Rating summary scores. The blue circles represent the mean RSCRs of Star-Rating summary score quartiles. The correlation between THA/TKA complications and Star-Rating summary score is -0.185, which suggests that hospitals with lower THA/TKA RSCRs are more likely to have higher Star-Rating summary scores especially at the extremes.

**Figure 1**

Figure 1 shows the Box-whisker plots of the THA/TKA complications measure RSCRs within each quartile of Star-Rating summary scores. The blue circles represent the mean RSCRs of Star-Rating summary score quartiles. The correlation between THA/TKA complications and Star-Rating summary score is -0.185, which suggests that hospitals with lower THA/TKA RSCRs are more likely to have higher Star-Rating summary scores.

Comparison to Hospital Surgical Admission Volume

Table 3 illustrates the relationship between deciles of admission volume and THA/TKA RSCRs. There is a general trend that high volume hospitals (those in the upper deciles) have lower RSCRs than hospitals in other volume deciles.

Table 3. Relationship Between Admission Volume and THA/TKA RSCRs

| - | **25≤N (N=2,763)** | | |
| --- | --- | --- | --- |
| **Correlation coefficient between admission volumes and RSCRs** | -0.25658  <.0001 | | |
| **Deciles of volume** | **# of Hospitals** | **Volume Range** | **Mean RSCR** |
| **0%~10%** | 283 | 25-43 | 2.51 |
| **10%~20%** | 274 | 44-67 | 2.54 |
| **20%~30%** | 273 | 68-100 | 2.54 |
| **30%~40%** | 276 | 101-144 | 2.62 |
| **40%~50%** | 276 | 145-200 | 2.55 |
| **50%~60%** | 274 | 201-278 | 2.55 |
| **60%~70%** | 275 | 279-380 | 2.49 |
| **70%~80%** | 280 | 382-528 | 2.36 |
| **80%~90%** | 276 | 529-804 | 2.35 |
| **90%~100%** | 276 | 808-9,018 | 2.10 |

Validation of Claims-Based Definition of Complications

Overall measure agreement was 93% (598/644 patients). More specifically, there were 598 patients who either had a complication coded in the claims and a complication was also documented in the medical record or who had no complication documented in both claims and medical record data. When we examined overall agreement in patients with and without complications, initial agreement was 86% for patients with a complication compared with 99% for patients without a complication. We proposed some minor changes to the measure on the basis of this validation study. Specifically, we determined that ICD-9 code 998.59, “Other postoperative infection,” was not sufficiently specific to sepsis, and the measure identified cases of sepsis that were not documented in the medical record. Therefore, we recommended removal of this code from the measure specifications. Secondly, we recommended combining wound infection and periprosthetic joint infection as a single complication in the measure specifications because these complications can be clinically difficult to differentiate. After the proposed measure changes were implemented, measure agreement between claims data and the medical record will increase to 99% (635/644 patients).

**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?*)

Empirical Validity Testing

This validation approach compares the THA/TKA complication measure results against the overall star rating scores. The Figure 1 Box Plot results demonstrate an observed trend of lower risk-standardized complications with higher star ratings, especially at the extremes, which supports measure score validity. Additionally, this validation approach compared various categories and deciles of hospital THA/TKA admission volume with THA/TKA complication measure scores in Table 3 – these results demonstrate an observed trend of higher hospital volume with lower complication measure scores. Overall, the results above show that the trend and direction of this association is in line with what would be expected.

Validation of Claims-Based Definition of Complications

The administrative claims-based and medical record data showed a high level of agreement in how they identified complications in the validity testing that was performed. There was overall measure agreement between the claims data and the medical record on the measure outcome in 99% of the cases after improving the claims-based definition of complication.

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

**2b2. EXCLUSIONS ANALYSIS**

**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*)

All exclusions were determined by careful clinical review and have been made based on clinically relevant decisions to ensure accurate calculation of the measure. To ascertain impact of exclusions on the cohort, we examined overall frequencies and proportions of the total cohort excluded for each exclusion criterion (**Testing Dataset**). These exclusions are consistent with similar NQF-endorsed outcome measures. Rationales for the exclusions are detailed in data field S.9 (Denominator Exclusions).

**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 **Testing Dataset (Table 4),** below is the distribution of exclusions among hospitals with 25 or more admissions:

**Table 4**. Frequency and Distribution of Exclusions Across Hospitals

| **Exclusion** | **N** | **%** | **Distribution across hospitals (N=2,789): Minimum, 25th percentile, 50th percentile, 75th percentile, maximum** |
| --- | --- | --- | --- |
| 1. Discharged against medical advice (AMA) | 156 | 0.02 | (0.0, 0.0, 0.0, 0.0, 2.94) |
| 2. Without at least 90 days post-discharge enrollment in FFS Medicare for index admissions | 7,815 | 0.77 | (0.0, 0.0, 0.60, 1.20, 8.57) |
| 3. Had more than two THA/TKA procedure codes during the index hospitalization | 1 | 0.00 | (0.0, 0.0, 0.0, 0.0, 0.09) |

**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*)

**Exclusion 1** (patients who are discharged AMA) accounts for 0.02% of all index admissions excluded from the initial index cohort. This exclusion is needed for acceptability of the measure to hospitals, who do not have the opportunity to adequately deliver full care. Because a very small percent of patients are excluded, this exclusion is unlikely to affect measure score.

**Exclusion 2** (patients without at least 90 days of post-discharge enrollment in FFS Medicare for index admissions) accounts for 0.77% of all index admissions excluded from the initial cohort. This exclusion is needed because the 90-day complication outcome cannot be assessed in this group since claims data are used to determine whether a patient has experienced complications. Because a very small percent of patients are excluded, this exclusion is unlikely to affect measure score.

**Exclusion 3** (patients with more than two THA/TKA procedure codes during the index hospitalization) accounts for only 1 of all index procedures excluded from the initial index cohort. Although clinically possible, it is highly unlikely that patients would receive more than two elective THA/TKA procedures in one hospitalization, which may reflect a coding error.

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

**2b3. 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***](#section2b5)***.***

**2b3.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** 33 **risk factors**

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

**Other,** Click here to enter description

**2b3.1.1 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.**

See risk model specifications in Section 2b3.4a and the attached data dictionary.

**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**.

N/A. This measure is risk adjusted.

**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?  
**Selecting Risk Variables**

Our goal in selecting risk factors for adjustment was to develop parsimonious models that included clinically relevant variables strongly associated with the risk of complication in the 90 days following an index procedure. We used a two stage approach, first identifying the comorbidity or clinical status risk factors that were most important in predicting the outcome, then considering the potential addition of social risk factors.

The original measure was developed with ICD-9. When ICD-10 became effective in 2015, we transitioned the measure to use ICD-10 codes as well. ICD-10 codes were identified using 2015 GEM mapping software. We then enlisted the help of clinicians with expertise in relevant areas to select and evaluate which ICD-10 codes map to the ICD-9 codes used to define this measure during development. A code set is attached in field S.2b. (Data Dictionary).

For risk model development, we started with Condition Categories (CCs) which are part of CMS’s Hierarchical Condition Categories (HCCs). The current HCC system groups the 70,000+ ICD-10-CM and 17,000+ ICD-9-CM codes into larger clinically coherent groups (201 CCs) that are used in models to predict mortality or other outcomes (Pope et al. 2001; 2011). The HCC system groups ICD- codes into larger groups that are used in models to predict medical care utilization, mortality, or other related measures.

To select candidate variables, a team of clinicians reviewed all CCs and excluded those that were not relevant to the Medicare population or that were not clinically relevant to the complication outcome (for example, attention deficit disorder, female infertility). All potentially clinically relevant CCs were included as candidate variables and, consistent with CMS’s other claims-based measures, some of those CCs were then combined into clinically coherent CC groupings.

To inform final variable selection, a modified approach to stepwise logistic regression was performed. The Development Sample was used to create 1,000 “bootstrap” samples. For each sample, we ran a logistic stepwise regression that included the candidate variables. The results (not shown in this report) were summarized to show the percentage of times that each of the candidate variables was significantly associated with mortality (p<0.01) in each of the 1,000 repeated samples (for example, 90 percent would mean that the candidate variable was selected as significant at p<0.01 in 90 percent of the times). We also assessed the direction and magnitude of the regression coefficients.

The clinical team reviewed these results and decided to retain risk adjustment variables above a predetermined cutoff, because they demonstrated a strong and stable association with risk of complication and were clinically relevant. Additionally, specific variables with particular clinical relevance to the risk of complications were forced into the model (regardless of percent selection) to ensure appropriate risk adjustment for THA/TKA. These included variables representing markers for end of life/frailty, such as:

Markers for end of life/frailty:

* Decubitus Ulcer or Chronic Skin Ulcer (CC 157-CC 161)
* Metastatic and Other Major Cancers (CC 8-CC 12)
* Osteoporosis and Other Bone/Cartilage Disorders (CC 43)
* Chronic Kidney Disease, Stage 5 (CC 136)
* Hemiplegia, Paraplegia, Paralysis, Functional disability (CC 70-CC 74, CC 103, CC 104, CC 189-CC 190)
* Stroke (CC 99-CC 100)

This resulted in a final risk-adjustment model that included 33 variables.

Social Risk Factors

We weigh SRF adjustment using a comprehensive approach that evaluates the following:

* Well-supported conceptual model for influence of SRFs on measure outcome (detailed below);
* Feasibility of testing meaningful SRFs in available data (section 1.8); and
* Empiric testing of SRFs (section 2b3.4b).

Below, we summarize the findings of the literature review and conceptual pathways by which social risk factors may influence risk of the outcome, as well as the statistical methods for SRF empiric testing. Our conceptualization of the pathways by which patients’ social risk factors affect the outcome is informed by the literature cited below and IMPACT Act–funded work by the National Academy of Science, Engineering and Medicine (NASEM) and the Department of Health and Human Services Assistant Secretary for Policy and Evaluation (ASPE).

Causal Pathways for Social Risk Variable Selection

Although some recent literature evaluates the relationship between patient SRFs and the complication outcome, few studies directly address causal pathways or examine the role of the hospital in these pathways (see, for example Gopaldas et al 2009; Kim et al., 2007; LaPar et al., 2010; 2012; Trivedi et al., 2014; Buntin et al., 2017; Borza et al., 2019). Moreover, the current literature examines a wide range of conditions and risk variables with no clear consensus on which risk factors demonstrate the strongest relationship with complication.

The social risk factors that have been examined in the literature can be categorized into three domains: (1) patient-level variables, (2) neighborhood/community-level variables, and (3) hospital-level variables.

Patient-level variables describe characteristics of individual patients, and include the patient’s income or education level. Neighborhood/community-level variables use information from sources such as the American Community Survey as either a proxy for individual patient-level data or to measure environmental factors. Studies using these variables use one dimensional measures such as median household income or composite measures such as the AHRQ-validated SES index score (Blum et al., 2014; Courtney et al., 2016; Martsolf et al., 2016; White et al., 2018). Some of these variables may include the local availability of clinical providers (Herrin et al., 2015; Herrin et al., 2016). Hospital-level variables measure attributes of the hospital which may be related to patient risk. Examples of hospital-level variables used in studies are ZIP code characteristics aggregated to the hospital level or the proportion of Medicaid patients served in the hospital (Gilman et al., 2014; Joynt et al., 2013; Jha et al., 2013; Xu et al., 2018).

The conceptual relationship, or potential causal pathways by which these possible social risk factors influence the risk of complication following an acute illness or major surgery, like the factors themselves, are varied and complex. There are at least four potential pathways that are important to consider:

1. **Patients with social risk factors may have worse health at the time of hospital admission**. Patients who have lower income/education/literacy or unstable housing may have a worse general health status and may present for their hospitalization or procedure with a greater severity of underlying illness. These social risk factors, which are characterized by patient-level or neighborhood/community-level (as proxy for patient-level) variables, may contribute to worse health status at admission due to competing priorities (restrictions based on job), lack of access to care (geographic, cultural, or financial), or lack of health insurance. Given that these risk factors all lead to worse general health status, this causal pathway should be largely accounted for by current clinical risk-adjustment.
2. **Patients with social risk factors often receive care at lower quality hospitals**. Patients of lower income, lower education, or unstable housing have inequitable access to high quality facilities, in part, because such facilities are less likely to be found in geographic areas with large populations of poor patients. Thus, patients with low income are more likely to be seen in lower quality hospitals, which can explain increased risk of complications following hospitalization.
3. **Patients with social risk factors may receive differential care within a hospital**. The third major pathway by which social risk factors may contribute to complications risk is that patients may not receive equivalent care within a facility. For example, patients with social risk factors such as lower education may require differentiated care (e.g. provision of lower literacy information – that they do not receive).
4. **Patients with social risk factors may experience worse health outcomes beyond the control of the health care system.** Some social risk factors, such as income or wealth, may affect the likelihood of complications without directly affecting health status at admission or the quality of care received during the hospital stay. For instance, while a hospital may make appropriate care decisions and provide tailored care and education, a lower-income patient may have a worse outcome post-discharge due to competing financial priorities which don’t allow for adequate recuperation or access to needed treatments, or a lack of access to care outside of the hospital.

Although we analytically aim to separate these pathways to the extent possible, we acknowledge that risk factors often act on multiple pathways, and as such, individual pathways can be complex to distinguish analytically. Further, some social risk factors, despite having a strong conceptual relationship with worse outcomes, may not have statistically meaningful effects on the risk model. They also have different implications on the decision to risk adjust or not.

Based on this model and the considerations outlined in section 1.8 - namely, that the AHRQ SES index and dual eligibility variables aim to capture the SRFs that are likely to influence these pathways (income, education, housing, and community factors) - the following social risk variables were considered for risk-adjustment:

* Dual eligible status
* AHRQ SES index

Statistical Methods

We assessed the relationship between the SRF variables with the outcome and examined the incremental effect in a multivariable model. For this measure, we also examined the extent to which the addition of any one of these variables improved model performance or changed hospital results.

One concern with including SRFs in a model is that their effect may be at either the patient or the hospital level. For example, low SES may increase the risk of complications because patients of low SES have an individual higher risk (patient-level effect) or because patients of low SES are more often admitted to hospitals with higher overall complication rates (hospital-level effect). Identifying the relative contribution of the hospital level is important in considering whether a factor should be included in risk adjustment; if an effect is primarily a hospital-level effect, adjusting for it is equivalent to adjusting for differences in hospital quality. Thus, as an additional step, we assessed whether there was a “contextual effect” at the hospital level. To do this, we performed a decomposition analysis to assess the independent effects of the SRF variables at the patient level and the hospital level. If, for example, the elevated risk of complications for patients of low SES were largely due to lower quality/higher complications risk in hospitals with more patients of low SES, then a significant hospital-level effect would be expected with little-to-no patient-level effect. However, if the increased complications risk were solely related to higher risk for patients of low SES regardless of hospital effect, then a significant patient-level effect would be expected and a significant hospital-level effect would not be expected.

Specifically, we modeled the SRF variables as follows, let Xij be a binary indicator of the SRF status of the ith patient at the jth hospital, and **X**j the percent of patients at hospital j with Xij = 1. Then we added both Xij ≡ Xpatient and Xj ≡ Xhospital  to the model. The first variable, Xpatient, represents the effect of the risk factor at the patient level (sometimes called the “within” hospital effect), and the second variable, Xhospital, represents the effect at the hospital level (sometimes called the “between” hospital effect). By including both of these in the same model, we can assess whether these are independent effects, whether one effect dominates the other, or whether only one of these effects contributes. This analysis allows us to simultaneously estimate the independent effects of: 1) hospitals with higher or lower proportions of low SES patients on the complication rate of an average patient; and 2) a patient’s SES on their own complication rates when seen at an average hospital.

It is very important to note, however, that even in the presence of a significant patient-level effect and absence of a significant hospital-level effect, the increased risk could be partly or entirely due to the quality of care patients receive in the hospital. For example, biased or differential care provided within a hospital to low-income patients as compared to high-income patients would exert its impact at the level of individual patients, and therefore be a patient-level effect.

It is also important to note that the patient-level and hospital-level coefficients cannot be quantitatively compared because the patient’s SES circumstance in the model is binary whereas the hospital’s proportion of low SES patients is continuous. Therefore, in order to quantitatively compare the relative size of the patient and hospital effects, we calculated a range of predicted probabilities of complications based on the fitted model.

Specifically, to estimate an average hospital effect, we calculated the predicted probabilities for the following scenarios: (1) Assuming all patients do not have the risk factor (Xij =0) and hospital level risk factor is at 5% percentile (P5) of all hospital values; (2) Assuming all patients do not have the risk factor and hospital level risk factor is at 95% percentile (P95); (3) Assuming all patients do have the risk factor (Xij =1) and hospital level risk factor is at 5% percentile (P5); (4) Assuming all patients have the risk factor and hospital level risk factor is at 95% percentile (P95). The average hospital effect is estimated by ((2)-(1) + (4)-(3))/2 (P95-P5). Then, to estimate an average patient effect, we first calculated the predicted probabilities by assuming patient-level risk factor equal to 0 or 1 at different hospital risk factor percentiles (0%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 100%). Then at each of those percentiles, we could obtain the difference of predicted probabilities between all patients not having the risk factor and then all patients having the risk factor. We calculated the average of those differences in predicted probabilities (‘delta’) as the patient effect.

In summary, the difference in predicted probabilities at the 95th and 5th percentiles (P95-P5) estimates the hospital-level effect of the SRF on complications. The difference in predicted probabilities when all patients have and do not have the SES risk factor (delta) estimates the patient-level effect of the SES risk factor on complications. The hospital-level effect is greater than the patient-level effect when P95-P5 is greater than delta. We used P95 and P5 rather than the maximum (P100) and minimum (P0) to avoid outlier values.

We also performed the same analysis for several clinical covariates to contrast the relative contributions of patient- and hospital-level effects of clinical variables to the relative contributions for the SRFs.

References

Blum AB, Egorova NN, Sosunov EA, et al. Impact of socioeconomic status measures on hospital profiling in New York City. Circulation Cardiovascular quality and outcomes 2014; 7:391-7.

Borza T, Oerline MK, Skolarus TA, et al. Association Between Hospital Participation in Medicare Shared Savings Program Accountable Care Organizations and Readmission Following Major Surgery. Ann Surg. 2019;269(5):873-878. doi:10.1097/SLA.0000000000002737.

Buntin MB, Ayanian JZ. Social Risk Factors and Equity in Medicare Payment. *New England Journal of Medicine.* 2017;376(6):507-510.

Committee on Accounting for Socioeconomic Status in Medicare Payment Programs; Board on Population Health and Public Health Practice; Board on Health Care Services; Institute of Medicine; National Academies of Sciences, Engineering, and Medicine. Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. Washington (DC): National Academies Press (US); 2016 Jan 12. (<https://www.ncbi.nlm.nih.gov/books/NBK338754/doi:10.17226/21858>)

Courtney M, Huddleston J, Iorio R, Markel D. Socioeconomic Risk Adjustment Models for Reimbursement Are Necessary in Primary Total Joint Arthroplasty. July 2016; 32(1):1-5. <https://doi.org/10.1016/j.arth.2016.06.050>.

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation. Report to Congress: Social Risk Factors and Performance under Medicare’s Value-based Payment Programs. December 21, 2016. (<https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs>).

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare’s Value-based Purchasing Programs. 2020; <https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf>. Accessed July 2, 2020.

Gilman M, Adams EK, Hockenberry JM, et al. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. Health Aff (Millwood). Aug 2014; 33(8):1314-22.

Gopaldas R R, Chu D., “Predictors of surgical mortality and discharge status after coronary artery bypass grafting in patients 80 years and older.” The American Journal of Surgery. 2009; 198(5): 633-638.

Herrin J, Kenward K, Joshi MS, Audet AM, Hines SJ. Assessing Community Quality of Health Care. Health Serv Res. 2016 Feb;51(1):98-116. doi: 10.1111/1475-6773.12322. Epub 2015 Jun 11. PMID: 26096649; PMCID: PMC4722214.

Herrin J, St Andre J, Kenward K, Joshi MS, Audet AM, Hines SC. Community factors and hospital readmission rates. Health Serv Res. 2015 Feb;50(1):20-39. doi: 10.1111/1475-6773.12177. Epub 2014 Apr 9. PMID: 24712374; PMCID: PMC4319869.

Jha AK, Orav EJ, Epstein AM. Low-quality, high-cost hospitals, mainly in South, care for sharply higher shares of elderly black, Hispanic, and medicaid patients. Health affairs 2011; 30:1904-11.

Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. JAMA. Jan 23 2013; 309(4):342-3.

Joynt KE, Orav EJ, Jha AK. Thirty-day readmission rates for Medicare beneficiaries by race and site of care. JAMA. 2011 Feb 16; 305(7):675-81. doi: 10.1001/jama.2011.123.

Kim C, Diez A V, Diez Roux T, Hofer P, Nallamothu B K, Bernstein S J, Rogers M, “Area socioeconomic status and mortality after coronary artery bypass graft surgery: The role of hospital volume.” Clinical Investigation Outcomes, Health Policy, and Managed Care. 2007; 154(2): 385-390.

Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes. Circulation. 2006; 113: 456-462. Available at: http://circ.ahajournals.org/content/113/3/456.full.pdf+html. Accessed January 14, 2016.

LaPar D J, Bhamidipati C M, et al. “Primary Payer Status Affects Mortality for Major Surgical Operations.” Annals of Surgery. 2010; 252(3): 544-551.

LaPar D J, Stukenborg G J, et al “Primary Payer Status Is Associated With Mortality and Resource Utilization for Coronary Artery Bypass Grafting.” Circulation. 2012; 126:132-139.

Martsolf G, Barrett M, Weiss A, Kandrack R, Washington R, Steiner C, Mehrotra A, SooHoo N, Coffey R. Impact of Race/Ethnicity and Socioeconomic Status on Risk-Adjusted Hospital Readmission Rates Following Hip and Knee Arthroplasty, The Journal of Bone and Joint Surgery. 2016;98(16):1385-1391. <https://doi.org/10.2106/JBJS.15.00884>.

Normand S-LT, Shahian DM. Statistical and Clinical Aspects of Hospital Outcomes Profiling. 2007/05 2007:206-226.

Pope GC, Ellis RP, Ash AS, et al. Diagnostic cost group hierarchical condition category models for Medicare risk adjustment. Final Report to the Health Care Financing Administration under Contract Number 500-95-048. 2000; <http://www.cms.hhs.gov/Reports/downloads/pope_2000_2.pdf>. Accessed February 25, 2020.

Pope GC, Kautter J, Ingber MJ, et al. Evaluation of the CMS-HCC Risk Adjustment Model: Final Report. 2011; <https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/downloads/evaluation_risk_adj_model_2011.pdf>. Accessed February 25, 2020.

Singh JA, Kallan MJ, Chen Y, Parks ML, Ibrahim SA. Association of Race/Ethnicity With Hospital Discharge Disposition After Elective Total Knee Arthroplasty. JAMA Network Open. 2019;2(10):e1914259. doi:10.1001/jamanetworkopen.2019.14259.

Trivedi AN, Nsa W, Hausmann LR, et al. Quality and equity of care in U.S. hospitals. The New England journal of medicine 2014; 371:2298-308.

White, R.S., Sastow, D.L., Gaber-Baylis, L.K. *et al.* Readmission Rates and Diagnoses Following Total Hip Replacement in Relation to Insurance Payer Status, Race and Ethnicity, and Income Status. *J. Racial and Ethnic Health Disparities* 5, 1202–1214 (2018). <https://doi.org/10.1007/s40615-018-0467-0>.

Xu HF, White RS, Sastow DL, Andreae MH, Gaber-Baylis LK, Turnbull ZA. Medicaid insurance as primary payer predicts increased mortality after total hip replacement in the state inpatient databases of California, Florida and New York. *J Clin Anesth*. 2017;43:24‐32. doi:10.1016/j.jclinane.2017.09.008.

**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?**

The table below shows the final variables in the model in the testing dataset with associated odds ratios (OR) and 95 percent confidence intervals (CI).

**Table 5**. Adjusted OR and 95% CIs for the THA/TKA Hierarchical Logistic Regression Model over Different Time Periods in the **Testing Dataset**

| **Variable** | **04/2016-03/2017 OR (95%)** | **04/2017-03/2018 OR (95%)** | **04/2018-03/2019 OR (95%)** | **04/2016-03/2019 OR (95%)** |
| --- | --- | --- | --- | --- |
| Age minus 65 (years above 65, continuous) | 1.03 (1.03-1.03) | 1.03 (1.03-1.04) | 1.03 (1.02-1.03) | 1.03 (1.03-1.03) |
| Male | 1.13 (1.08-1.18) | 1.15 (1.10-1.21) | 1.17 (1.11-1.23) | 1.15 (1.12-1.18) |
| Index admissions with an elective THA procedure | 1.27 (1.22-1.33) | 1.30 (1.24-1.36) | 1.26 (1.19-1.32) | 1.28 (1.24-1.31) |
| Number of procedures (two vs. one) | 1.70 (1.47-1.96) | 1.69 (1.45-1.97) | 1.65 (1.39-1.96) | 1.69 (1.54-1.85) |
| Other congenital deformity of hip (joint) | 1.66 (1.17-2.38) | 1.89 (1.35-2.65) | 1.25 (0.82-1.90) | 1.60 (1.30-1.98) |
| Post traumatic osteoarthritis | 1.10 (0.93-1.31) | 1.13 (0.95-1.35) | 1.06 (0.87-1.30) | 1.10 (0.99-1.23) |
| Metastatic cancer and acute leukemia (CC 8) | 1.04 (0.81-1.34) | 0.93 (0.71-1.23) | 0.91 (0.68-1.21) | 0.96  (0.82-1.12) |
| Other major cancers (CC 9-12) | 0.96 (0.90-1.02) | 0.96 (0.90-1.03) | 0.89  (0.82-0.96) | 0.94 (0.90-0.98) |
| Respiratory/heart/digestive/urinary/other neoplasms (CC 13-15) | 1.00 (0.95-1.06) | 0.94 (0.89-1.00) | 0.93  (0.87-0.99) | 0.96 (0.93-0.99) |
| Diabetes mellitus (DM) or DM complications (CC 17-19, 122-123) | 1.13 (1.08-1.19) | 1.12 (1.07-1.18) | 1.10 (1.04-1.16) | 1.12 (1.08-1.15) |
| Protein-calorie malnutrition (CC 21) | 2.51 (2.15-2.92) | 1.69 (1.43-2.00) | 1.78 (1.49-2.12) | 1.97 (1.79-2.16) |
| Morbid obesity (CC 22) | 1.65 (1.54-1.76) | 1.60 (1.50-1.71) | 1.64 (1.53-1.76) | 1.63 (1.56-1.69) |
| Bone/joint/muscle infections/necrosis (CC 39) | 1.10 (0.98-1.23) | 1.21 (1.08-1.35) | 1.40 (1.25-1.57) | 1.22 (1.15-1.31) |
| Rheumatoid arthritis and inflammatory connective tissue disease (CC 40) | 1.14 (1.06-1.22) | 1.18 (1.10-1.26) | 1.23 (1.14-1.32) | 1.18 (1.13-1.23) |
| Osteoarthritis of hip or knee (CC 42) | 0.97 (0.85-1.10) | 0.94 (0.83-1.08) | 0.97 (0.83-1.12) | 0.96 (0.89-1.04) |
| Osteoporosis and other bone/cartilage disorders (CC 43) | 1.03 (0.98-1.09) | 1.03 (0.98-1.09) | 1.04 (0.98-1.10) | 1.03 (1.00-1.07) |
| Dementia or other specified brain disorders (CC 51-53) | 1.20 (1.10-1.31) | 1.32 (1.21-1.45) | 1.17 (1.06-1.30) | 1.23 (1.17-1.30) |
| Major psychiatric disorders (CC 57-59) | 1.43 (1.31-1.56) | 1.41 (1.29-1.53) | 1.36 (1.25-1.49) | 1.39 (1.32-1.46) |
| Hemiplegia, paraplegia, paralysis, functional disability (CC 70-74, 103-104, 189-190) | 1.22 (1.05-1.42) | 1.17 (1.00-1.36) | 1.31 (1.12-1.53) | 1.23 (1.12-1.34) |
| Cardio-respiratory failure and shock (CC 84 plus ICD-10-CM codes R09.01 and R09.02, for discharges on or after October 1, 2015; CC 84 plus ICD-9-CM codes 799.01 and 799.02, for discharges prior to October 1, 2015) | 1.24 (1.12-1.38) | 1.24 (1.11-1.38) | 1.39 (1.25-1.55) | 1.28 (1.20-1.36) |
| Coronary atherosclerosis or angina (CC 88-89) | 1.33 (1.27-1.40) | 1.28 (1.22-1.35) | 1.29 (1.22-1.36) | 1.30 (1.26-1.34) |
| Stroke (CC 99-100) | 1.01 (0.88-1.16) | 1.19 (1.04-1.36) | 1.12 (0.97-1.29) | 1.10 (1.02-1.19) |
| Vascular or circulatory disease (CC 106-109) | 1.14 (1.08-1.20) | 1.11 (1.05-1.17) | 1.16 (1.09-1.22) | 1.13 (1.10-1.17) |
| Chronic obstructive pulmonary disease (COPD) (CC 111) | 1.5 8(1.49-1.67) | 1.63 (1.54-1.73) | 1.55 (1.45-1.65) | 1.58 (1.53-1.64) |
| Pneumonia (CC 114-116) | 1.17 (1.07-1.29) | 1.15 (1.04-1.27) | 1.14 (1.03-1.27) | 1.16 (1.09-1.22) |
| Pleural effusion/pneumothorax (CC 117) | 0.99 (0.86-1.14) | 0.98 (0.85-1.13) | 1.00 (0.86-1.17) | 0.99 (0.91-1.08) |
| Dialysis status (CC 134) | 1.14 (0.83-1.57) | 1.91 (1.45-2.53) | 1.38 (1.00-1.91) | 1.46 (1.22-1.74) |
| Renal failure (CC 135-140) | 1.33 (1.26-1.41) | 1.35 (1.28-1.43) | 1.27 (1.20-1.35) | 1.31 (1.27-1.36) |
| Decubitus ulcer or chronic skin ulcer (CC 157-161) | 1.27 (1.13-1.43) | 1.28 (1.14-1.44) | 1.34 (1.18-1.51) | 1.30 (1.21-1.39) |
| Trauma (CC 166-168, 170-173) | 1.16 (1.06-1.27) | 1.15 (1.04-1.26) | 1.13 (1.02-1.25) | 1.14 (1.08-1.21) |
| Vertebral fractures without spinal cord injury (CC 169) | 1.03 (0.86-1.23) | 0.95 (0.78-1.15) | 1.11 (0.92-1.34) | 1.03 (0.92-1.15) |
| Other injuries (modified) (CC 174) | 1.13 (1.07-1.19) | 1.10 (1.05-1.16) | 1.12 (1.06-1.18) | 1.12 (1.08-1.15) |
| Major complications of medical care and trauma (CC 176-177) | 1.23 (1.13-1.34) | 1.34 (1.23-1.46) | 1.20 (1.09-1.32) | 1.26 (1.19-1.32) |

**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.**

Throughout this section, we present new SRF testing results based on the current testing dataset (2020); in addition, we show prior analyses included in the 2016 endorsement maintenance forms for comparison purposes.

Variation in prevalence of the factor across measured entities in 2020 and 2016 (Table 6)

| **SRFs** | **2020 Prevalence % (IQR)** | **2016 Prevalence % (IQR)** |
| --- | --- | --- |
| Dual | 3.40% (1.4-7.8%) | 6.70% (3.9-11.7%) |
| AHRQ Low SES | 11.7% (5.0-23.9%) | 12.9% (6.4-24.0%) |

The prevalence of social risk factors in the THA/TKA cohort varies widely across measured entities in 2020. The median percentage of dual eligible patients was 3.40% (IQR 1.4-7.8%) and the median percentage of patients with an AHRQ SES index score adjusted for cost of living at the census block group level equal to or below 42.7 (lowest quartile) was 11.7% (IQR 5.0-23.9%) in 2020. These results are consistent with the 2016 results presented above. Overall, there has been a decline in dually eligible and AHRQ Low SES patients since last endorsement maintenance.

Comparison of observed complication rates in patients with and without social risk in 2020 and 2016 (Table 7)

|  |  |  |
| --- | --- | --- |
| **SRFs** | **2020 Observed Rate** | **2016 Observed Rate** |
| Dual (vs. Non-Dual) | 3.7% (vs. 2.4%) | 4.3% (vs. 3.1%) |
| AHRQ Low SES (vs. SES score above 42.7) | 2.9% (vs. 2.4%) | 3.5% (vs. 3.1%) |

The patient-level observed THA/TKA complication rates are higher for dual-eligible patients (3.7%) compared with 2.4% for non-dual patients in 2020. Similarly, the complication rate for patients with an AHRQ SES index score equal to or below 42.7 was 2.9% compared with 2.4% for patients with an AHRQ SES index score above 42.7 in 2020. For both SRF variables, patient-level complication rates have declined among all characteristic groups of patients.

Incremental effect of SRF variables in a multivariable model in 2020 and 2016

We examined the strength and significance of the SRF variables in the context of a multivariable model. When we include these variables in a multivariable model that includes all of the claims-based clinical variables, the effect size of each of these variables is moderate. In 2020, dual eligibility and the AHRQ SES index have effect sizes (odds ratios) of 1.26 and 1.16 when added separately to the model, with a slightly larger effect size than the 2016 findings (1.21 and 1.07, respectively). Furthermore, the effect size of each variable is slightly attenuated (1.24 and 1.13 for dual and SES) when both are added to the model simultaneously.

We also find that the c-statistic is essentially unchanged with the addition of any of these variables into the model (Table 8).

**Table 8**

| **THA/TKA Complications Models** | **2020 C-Statistic** | **2016 C-Statistic** |
| --- | --- | --- |
| Base Model: risk-adjusted model using the original clinical risk variables selected for the 2020 CMS public report of the THA/TKA complications measure | 0.65 | 0.65 |
| Base Model plus AHRQ Low SES based on beneficiary residential 9-digit ZIP codes (SES9) as a social risk variable | 0.65 | 0.65 |
| Base Model plus dual eligibility (dual) as a social risk variable | 0.65 | 0.65 |
| Base Model plus SES9 and dual as social risk variables | 0.65 | - |

We find that the addition of any of these variables into the model has little to no effect on hospital performance. We examined the change in hospitals’ RSCRs with the addition of any of these variables. The median absolute change in hospitals’ RSCRs when adding a dual eligibility indicator is 0.005% (interquartile range [IQR] -0.004% – 0.007%) with a correlation coefficient between RSCRs for each hospital with and without dual eligibility close to 1.000. The median absolute change in hospitals’ RSCRs when adding a low AHRQ SES Index score indicator to the model is 0.031% (IQR -0.006% – 0.041%) with a correlation coefficient between RSCRs for each hospital with and without an indicator for a low AHRQ SES Index score adjusted for cost of living at the census block group level is 0.982.

Contextual Effect Analysis

As described in 2b3.3a, we performed a decomposition analysis in 2020 and 2016 for each SRF variable to assess whether there was a corresponding contextual effect. In order to better interpret the magnitude of results, we performed the same analysis for selected clinical risk factors. The results are described in the tables/figures below.

Both the patient-level and hospital-level dual eligibility, and low AHRQ SES Index effects were significantly associated with THA/TKA complications in the decomposition analysis. **That the hospital level effects were significant indicates that if the dual eligible or low AHRQ SES Index variables were used in the model to adjust for patient-level differences, then some of the differences between hospitals would also be adjusted for, potentially obscuring a signal of hospital quality**.

To assess the relative contributions of the patient- and hospital-level effects, we calculated a range of predicted probabilities of complications for the SRF variables and clinical covariates (comorbidities), as described in section 2b3.3a. The results are presented in the figures and table below (table of predicted probabilities for SRF variables).

For the AHRQ SES index, the hospital-level effect (P95-P5) is greater than the patient-level effect (delta) (Figures 2 and 3; predicted probabilities for SRF variables); however, the patient-level effect (P95-P5) is greater than the hospital-level effect (delta) for dual-eligibility. For clinical variables, the patient-level effect (delta) is greater than the hospital-level effect (P95-P5) for renal failure and COPD (Figures 2 and 3; predicted probabilities for clinical variables). In sum, including SRF variables into the model would predominantly adjust for a hospital-level effect, which is an important signal of hospital quality.

In the context of our conceptual model, we find clear evidence supporting the first two mechanisms by which SRFs might be related to poor outcomes. First, we find that although unadjusted rates of complications are higher for patients of low SES, the addition of SRFs to the complications risk model, which already adjusts for clinical factors, makes very little difference. In particular, there is little-to-no change in model performance or hospital results with the addition of SRFs. This suggests that the model already largely accounts for the differences in clinical risk factors (degree of illness and comorbidities) among patients of varied SES.

Second, the predominance of the hospital-level effect of SRF variables in the decomposition analyses for 2020 and 2016 (Figures 2 and 3 below) suggests the risk associated with low SES is in large part due to lower quality of care at hospitals where more patients with these risk factors are treated; hospitals caring for socially- and economically-disadvantaged patients have higher complications risk for all of their patients. Patients with SRFs tend to receive care more frequently at lower quality hospitals compared with patients with high SRF indicators. Direct adjustment for patient SRFs would essentially “over adjust” the measure, that is to say, it would be adjusting for an endogenous factor, one that influences the outcome through the site of treatment (hospital), as much as through an attribute of the patient.

In comparison, we did not observe the same predominance of the hospital-level effect among the clinical covariates, reinforcing the sense that SRFs have a distinct causal pathway in their impact on complications risk.

**Table 9**. Parameter Estimates for Hospital-Level and Patient-Level in 2020 and 2016 from Decomposition Analysis

| **Parameter** | **2020 Estimate (standard error), p-value** | **2016 Estimate (standard error), p-value** |
| --- | --- | --- |
| Low SES census block group (AHRQ SES index linked to 9-digit ZIP – Adjusted for Cost of Living) – Patient Level | 0.074 (0.021), <0.001 | 0.045 (0.018), 0.0133 |
| Low SES census block group (AHRQ SES index linked to 9-digit ZIP – Adjusted for Cost of Living) – Hospital Level | 0.732 (0.082), <.0001 | 0.358 (0.069), <0.0001 |
| Dual-Eligible – Patient Level | 0.176 (0.031), <.0001 | 0.163 (0.023), <0.0001 |
| Dual-Eligible – Hospital Level | 0.723 (0.119), <.0001 | 0.507 (0.090), <0.0001 |

Figure 2. Decomposition Analysis for 2020, THA/TKA Complications

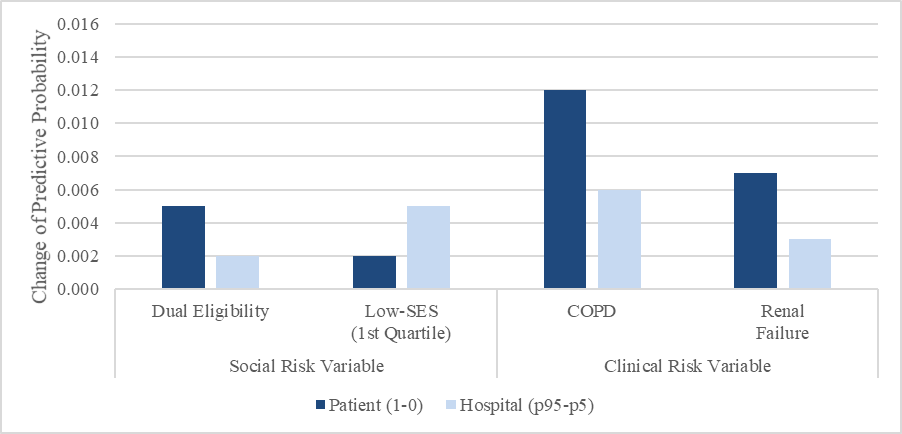
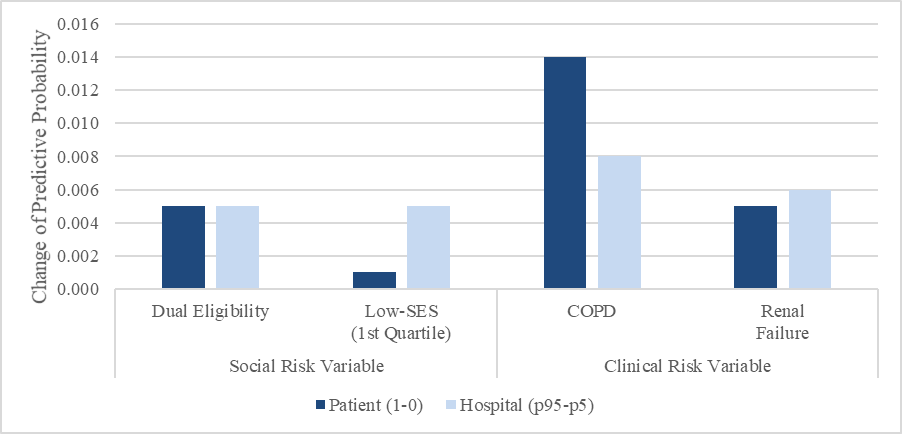
****

Figure 3. Decomposition Analysis for 2016, THA/TKA Complications

****

Summary

For risk-adjusted outcome measures, CMS first considers adjustment for clinical comorbidities, frailty indicators, and then examines additional risk imparted by SRFs after the potential for greater disease burden is included in the risk model (see section 2b3.3a). We believe that this is consistent with NQF current guidance and is appropriate given the evidence cited in our submission that people who experience greater social risk are more likely to have more disease burden compared with those who have less social risk; and that this is clearly not a signal of hospital quality. In addition, according to NQF guidance, developers should assess social risk factors for their contribution of unique variation in the outcome – that they are not redundant (NQF, 2014). Therefore, if clinical risk factors explain all or most of the patient variation in the outcome, then NQF guidance does not support adding social risk factors that account for relatively little variation. CMS’s decisions about which risk factors should be included in each measure’s risk-adjustment model are based on whether inclusion of such variables is likely to make the measures more successful at illuminating quality differences and motivating quality improvement. (This aim should be distinguished from decisions made in response to concerns about the impact of related payment programs on safety-net hospitals; concerns which can be addressed through other policy mechanisms.)

We found wide variation in the prevalence of the two SRFs we examined, with a large proportion of hospitals treating zero patients with these SRFs. We also found that both had some association with complication risk. However, adjustment for these factors did not have a material impact on hospital RSCRs, suggesting that existing clinical risk factors capture much of the risk related to social risk.

Ongoing research aims to identify valid patient-level social risk factors and highlight disparities related to social risk – in fact, ASPE’s latest report to Congress highlights which SRFs are valid in claims data, and that adjustment for SRFs in publicly reported quality measures is not recommended because providers should be accountable for overall outcomes, regardless of social risk (ASPE 2020). As additional variables become available, they will be considered for testing and inclusion within the measure. There are alternative ways that CMS considers adjusting for social risk as part of measure program implementation, such as stratification or peer grouping. CMS also considers confidentially reporting measure disparities to hospitals so that they have more detailed, actionable information about their patient population’s social risk. Given these empiric findings and program considerations, CMS chose not to include these two SRFs in the final risk model at this time.

We acknowledge the importance of balancing these competing considerations and are committed to constant refinement and improvement of risk adjustment models used in all measures. We will continue to reevaluate this model and available risk factors on an ongoing basis, with the goal of producing the most accurate and fair risk adjustment models for assessing provider performance.

References:

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare’s Value-based Purchasing Programs. 2020; <https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf>. Accessed July 2, 2020.

National Quality Forum (NQF). Risk adjustment for socioeconomic status or other sociodemographic factors: Technical report. 2014; <http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx>. Accessed June 16, 2020.

**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*)

Approach to assessing model performance

We computed three summary statistics for assessing model performance (Harrell and Shih, 2001) for the expanded cohort:

***Discrimination Statistics***

(1) Area under the receiver operating characteristic (ROC) curve (the c-statistic) is the probability that predicting the outcome is better than chance, which is a measure of how accurately a statistical model is able to distinguish between a patient with and without an outcome)

(2) Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects; therefore, we would hope to see a wide range between the lowest decile and highest decile

***Calibration Statistics***

(3) Over-fitting indices (over-fitting refers to the phenomenon in which a model accurately describes the relationship between predictive variables and outcome in the development dataset but fails to provide valid predictions in new patients)

We tested the performance of the model for **the development dataset** described in section 1.7.

References:

Harrell FE and Shih YC, Using full probability models to compute probabilities of actual interest to decision makers, *Int. J. Technol. Assess. Health Care* **17** (2001), pp. 17–26.

*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*)**:**

Development and Validation Dataset

First half of randomly split development sample: C-statistic = 0.69; Predictive ability (lowest decile %, highest decile %) = (2, 15)

Second half of randomly split development sample: C-statistic = 0.70; Predictive ability (lowest decile %, highest decile %) = (2, 15)

Results for the Testing Dataset

C-statistic = 0.65

Predictive ability (lowest decile %, highest decile %): (1.1, 5.9)

For comparison of model with and without inclusion of social risk factors, see above section.

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

For the original measure development cohort, the results are summarized below:

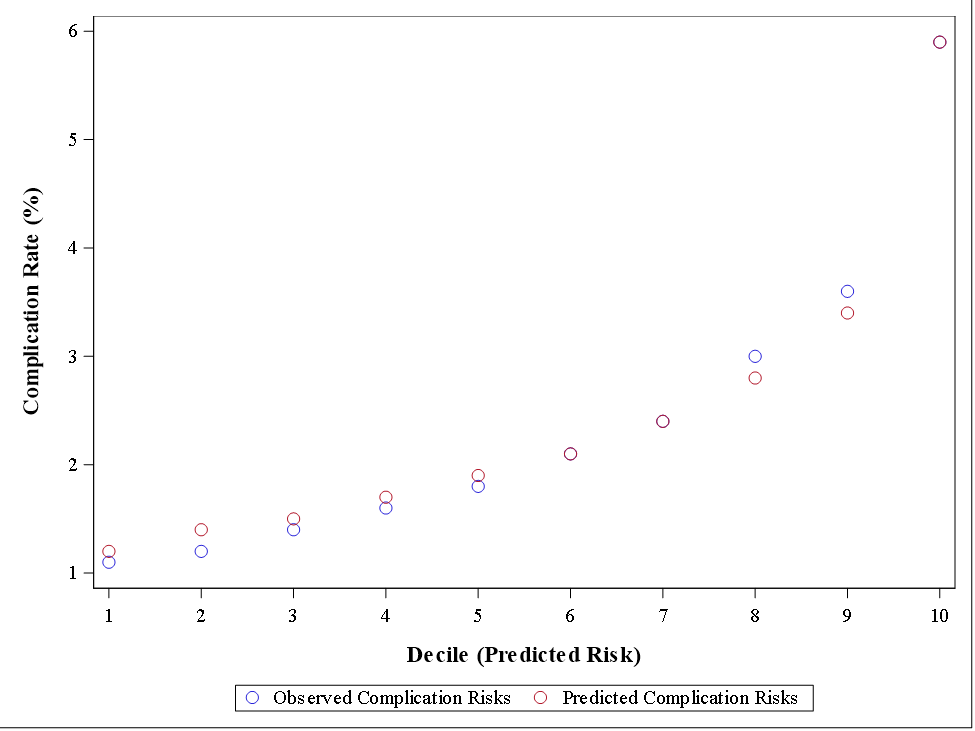
First half of split sample: Calibration: (0, 1)

Second half of split sample: Calibration: (0.04, 1.02)

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

The risk decile plot is a graphical depiction of the deciles calculated to measure predictive ability. Below, we present the risk decile plot showing the distributions for Medicare FFS data from April 2016 – March 2019 (Testing Dataset).

**Figure 4. Risk Decile Plot**



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

N/A

**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*)

***Discrimination Statistics***

The c-statistic of 0.65 indicate fair model discrimination. The model indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

***Calibration Statistics***

*Over-fitting (Calibration γ0, γ1)*

If the γ0 in the validation samples are substantially far from zero and the γ1 is substantially far from one, there is potential evidence of over-fitting. The calibration value of close to 0 at one end and close to 1 to the other end indicates calibration of the model.

***Risk Decile Plots***

Higher deciles of the predicted outcomes are associated with higher observed outcomes, which show a good calibration of the model. This plot indicates good discrimination of the model and good predictive ability.

***Overall Interpretation***

Interpreted together, our diagnostic results demonstrate the risk-adjustment model adequately controls for differences in patient characteristics (case mix) and is comparable to other outcome measures.

**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*)

N/A

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

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

**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)*

The measure score is hospital-specific risk-standardized complication rates. These rates are obtained as the ratio of predicted to expected complications, multiplied by the national unadjusted rate. The “predicted” number of complications (the numerator) is calculated using the coefficients estimated by regressing the risk factors and the hospital-specific intercept on the risk of complications. The estimated hospital-specific intercept is added to the sum of the estimated regression coefficients multiplied by the patient characteristics. The results are then transformed and summed over all patients attributed to a hospital to get a predicted value. The “expected” number of complications (the denominator) is obtained in the same manner, but a common intercept using all hospitals in our sample is added in place of the hospital-specific intercept. The results are then transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimated the model coefficients using the years of data in that period  
We characterize the degree of variability by:

1. Reporting the distribution of RSCRs:

For public reporting of the measure, CMS characterizes the uncertainty associated with the RSCR by estimating the 95% interval estimate. This is similar to a 95% confidence interval but is calculated differently. If the RSCR’s interval estimate does not include the national observed complication rate (because it is lower or higher than the rate), then CMS is confident that the hospital’s RSCR is different from the national rate, and describes the hospital on the Hospital Compare website as “better than the U.S. national rate” or “worse than the U.S. national rate.” If the interval includes the national rate, then CMS describes the hospital’s RSCR as “no different than the U.S. national rate” or “the difference is uncertain.” CMS does not classify performance for hospitals that have fewer than 25 cases in the three-year period.

1. Providing the median odds ratio (MOR) (Merlo et al, 2006). The median odds ratio represents the median increase in the odds of a complication within 30 days of a THA/TKA admission date on a single patient if the admission occurred at a higher risk hospital compared to a lower risk hospital. MOR quantifies the between-hospital variance in terms of odds ratio, it is comparable to the fixed effects odds ratio.

Reference:

Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Råstam L, Larsen K. (2006) A brief conceptual tutorial of multilevel analysis in social epidemiology: Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. J Epidemiol Community Health, 60(4):290-7.

**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*)

Analyses of Medicare FFS data show substantial variation in RSCRs among hospitals.

Figure 5. Distribution (Histogram) Of Hospital-Level THA/TKA RSCRs

Out of 3,418 hospitals in the measure cohort, 60 performed “better than the U.S. national rate,” 2,653 performed “no different from the U.S. national rate,” and 50 performed “worse than the U.S. national rate.” 655 were classified as “number of cases too small” (fewer than 25) to reliably tell how well the hospital is performing.

Out of 3,418 hospitals in the measure cohort, 60 performed “better than the U.S. national rate,” 2,653 performed “no different from the U.S. national rate,” and 50 performed “worse than the U.S. national rate.” 655 were classified as “number of cases too small” (fewer than 25) to reliably tell how well the hospital is performing.

The median odds ratio was 1.38.

**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?*)

The median odds ratio suggests a meaningful increase in the risk of complications if a patient has a THA/TKA procedure at a higher risk hospital compared to a lower risk hospital. A value of 1.38 indicates that a patient has a 38% increase in the odds of a complications at a higher risk performance hospital compared to a lower risk hospital, indicating the impact of quality on the outcome rate.

The variation in rates and number of performance outliers suggests there remain differences in the quality of care received across hospitals for THA/TKA procedures. This evidence supports continued measurement to reduce the variation.

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

**2b5. 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 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*)

N/A

**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*)

N/A

**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*)

N/A

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

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

**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*)

The THA/TKA complications measure used claims-based data for development and testing. There was no missing data in the development and testing data.

**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*)

N/A

**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*)

N/A