**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

*Last Updated 1/19/16*

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

**Measure Title**: Hybrid Hospital-Wide Readmission Measure with Claims and Electronic Health Record Data

**Date of Submission**: 1/29/2016

**Type of Measure:**

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

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

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) 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.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** 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.23*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Electronically abstracted from EHRs |

**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 inpatient claims, as well as electronically and manually abstracted electronic health record (EHR) data from several health systems. Data set 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.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

**1.5. 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*)

|  |
| --- |
| The number of measured entities 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*)

|  |
| --- |
| 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**.

|  |
| --- |
| Please note this model was developed using electronically extracted EHR data and merged inpatient claims data.  The datasets, dates, number of measured entities, and number of admissions used in each type of testing are as follows:  Measure Development and Testing:  For measure development and testing, we used a three-year dataset (**Dataset 1**) provided by Kaiser Permanente of Northern California. The dataset contained merged inpatient claims with clinical data elements derived from patients’ electronic health records (EHRs). This health system uses an Epic EHR system. The merged data were provided for all patients discharged from any of their 21 acute care hospitals from January 1, 2010 through December 31, 2012. We randomly split the first 2-years of this dataset (January 1, 2010 – December 31, 2011) into a “development sample” (used to develop a risk model) and a “validation sample” (used to re-test the model); the random split was stratified by hospital and the measure’s five specialty cohorts used to calculate the measure score. We re-tested the five risk models that make up the measure in the third year of data, from January 1, 2012 through December 31, 2012. This “2012 sample,” was used to look for temporal stability in the models’ performance.  In **Dataset 1**:  Number of admissions = 381,980  Number of hospitals = 21  Patient Descriptive Characteristics: mean age = 58 years; standard deviation = 21 years  %female = 62.6  The number of index admissions is listed below by specialty cohort.  Surgery:   * Development sample: 23,201 admissions * Validation sample: 23,490 admissions * 2012 sample: 25,471 admissions   Cardiorespiratory   * Development sample: 9,261 admissions * Validation sample: 9,364 admissions * 2012 sample: 9,070 admissions   Cardiovascular   * Development sample: 8,108 admissions * Validation sample: 8,037 admissions * 2012 sample: 8,338 admissions   Neurology   * Development sample: 4,400 * Validation sample: 4,348 * 2012 sample: 4,487   Medicine   * Development sample: 34,619 * Validation sample: 34,574 * 2012 sample: 35,747   For testing data Element and Measure Reliability Testing (Section 2a2)  **Dataset 1**  Validity Testing (Section 2b2)  **Dataset 1** was used for measure validity testing.  Three datasets were used to assess the feasibility or validity of the clinical data elements used in the measure’s risk models:  **Dataset 1:** (data element feasibility testing)  **Dataset 2**: (data element feasibility and validity testing)  Electronically extracted clinical data from three hospitals that used Cerner as their clinical EHR.   * Feasibility testing: 3 hospitals with 25,829; 56,812; and 29,586 admissions * Validity testing: 1 hospital with data abstracted from 368 admissions (subset of admissions above)   **Dataset 3**: (data element validity testing)  Data were electronically extracted from 1 hospital that used GE Centricity as their clinical EHR   * Validity testing: 1 hospital with data abstracted from 391 admissions   For testing of measure exclusions (Section 2b3)  **Dataset 1**  For testing of measure risk adjustment (Section 2b4)  **Dataset 1**  For testing to identify meaningful differences in performance (Section 2b5)  **Dataset 1**  For testing of socioeconomic status (SES) factors and race in risk models (Section 2b4)  **Dataset 4** and **Dataset 5** (Section 2b4)  The impact of socioeconomic factors was not directly tested in the Hybrid HWR measure due to lack of availability of EHR data from a nationally representative set of hospitals with patients who represent the full spectrum of socioeconomic status. Instead, we report results of testing done in the claims-only HWR measure.  **Dataset 4**: (2015 public reporting cohort version 4.0): Medicare Part A Inpatient Claims and Medicare Enrollment Database  Dates of Data: July 1, 2013 – June 30, 2014  Number of index admissions: 6,843,808  Number of hospitals: 4,772  Average age of patients: 78.3  We examined disparities in performance according to the proportion of patients in each hospital who were of African-American race and the proportion who were dual eligible for both Medicare and Medicaid insurances. We also used the AHRQ SES index score to study the association between performance measures and SES.  **Dataset 5**: The American Community Survey (2008-2012)  We also used the Agency for Healthcare Research and Quality (AHRQ) SES index score derived from the American Community Survey (2008-2012) to study the association between performance measures and socioeconomic status.  Data Elements  • African-American race and dual eligible status (i.e., enrolled in both Medicare and Medicaid) patient-level data are obtained from CMS enrollment data (**Dataset 4**)  • Validated AHRQ SES index score is a composite of 7 different variables found in the census data (the American Community Survey [2008-2012]) (**Dataset 5**) |

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

The impact of socioeconomic factors was not directly tested in the Hybrid HWR measure due to lack of availability of EHR data from a nationally representative set of hospitals with patients who represent the full spectrum of socioeconomic status. Instead, we report results of testing done in the claims-only HWR measure (**Datasets 4** and **5**).

Sociodemographic status incorporates socioeconomic variables as well as race into a more concise term. However, given the fact that socioeconomic risk factors are distinct from race and should be interpreted differently, we have decided to keep “socioeconomic status” and “race” as separate terms.

We selected SES and race variables to analyze after reviewing the literature and examining available national data sources. There is a large body of literature linking various SES factors and African-American race to worse health status and higher readmission risk (Blum et al., 2014; Eapen et al. 2015; Gilman et al., 2014; Hu et al., 2014; Joynt and Jha, 2013). Income, education, and occupational level are the most commonly examined variables. However, while literature directly examining how different SES factors or race might influence the likelihood of older, insured, Medicare patients of being readmitted within 30 days of an admission across multiple conditions is more limited, available studies suggest a consistent association between SES/race variables and risk of readmission. (Aseltine et al., 2015; Gu et al., 2014; Arbaje et al., 2008). The causal pathways for SES and race variable selection are described below in Section 2b4.3.

The SES and race variables used for analysis were:

* Dual eligible status (**Dataset 4**)
* African-American race (**Dataset 4**)
* AHRQ-validated SES index score (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) (**Dataset 5**)

In selecting variables, our intent was to be responsive to the NQF guidelines for measure developers in the context of the SDS Trial Period. Our approach has been to examine all patient-level indicators of both SES and race or ethnicity that are reliably available for all Medicare beneficiaries and linkable to claims data and to select those that are most valid.

Previous studies examining the validity of data on patients' race and ethnicity collected by CMS have shown that only the data identifying African-American beneficiaries have adequate sensitivity and specificity to be applied broadly in research or measures of quality. While using this variable is not ideal because it groups all non-African-American beneficiaries together, it is currently the only race variable available on all beneficiaries across the nation that is linkable to claims data.

We similarly 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 both our race and the dual-eligible variables, there is a body of literature demonstrating differential health care and health outcomes among beneficiaries indicating that these variables, while not ideal, also allow us to examine some of the pathways of interest.

Finally, we selected the AHRQ-validated SES index score because it is a well-validated and widely-used variable that describes the average socioeconomic status of people living in defined geographic areas. 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. Currently, the individual data elements used to calculate the score are available at the 5-digit zip code and census block levels only. The data are not currently available at the 9-digit zip code level. In this submission, we present analysis using the 5-digit level. However, we are currently performing analysis at the census block level, the most granular level possible. We hope to present the results of the census block-level analysis to the committee.

References:

Arbaje AI, Wolff JL, Yu Q, Powe NR, Anderson GF, Boult C. Post discharge environmental and socioeconomic factors and the likelihood of early hospital readmission among community-dwelling Medicare beneficiaries. The Gerontologist. 2008;48(4):495-504.

Aseltine RH, Jr., Yan J, Gruss CB, Wagner C, Katz M. Connecticut Hospital Readmissions Related to Chest Pain and Heart Failure: Differences by Race, Ethnicity, and Payer. Connecticut medicine. 2015;79(2):69-76.

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. May 2014; 7(3):391-397.

Eapen ZJ, McCoy LA, Fonarow GC, Yancy CW, Miranda ML, Peterson ED, Califf RM, HernandezAF. Utility of socioeconomic status in predicting 30-day outcomes after heart failure hospitalization. Circ Heart Fail. May 2015; 8(3):473-80.

Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. 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.

Gu Q, Koenig L, Faerberg J, Steinberg CR, Vaz C, Wheatley MP. The Medicare Hospital Readmissions Reduction Program: potential unintended consequences for hospitals serving vulnerable populations. Health services research. 2014;49(3):818-837.

Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health affairs (Project Hope). 2014; 33(5):778-785.

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

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

**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*)

|  |
| --- |
| Data Element Reliability  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 hospitals or 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. For example, “discharge disposition” is a variable in Medicare claims data that is not thought to be a reliable variable for identifying a transfer between two acute care facilities. Thus, we derive a variable using admission and discharge dates as a surrogate for “discharge disposition” to identify hospital admissions involving transfers. This allows us to identify these admissions using variables in the claims data which have greater reliability than the “discharge disposition” variable.  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.  We also assess the reliability of the claims data elements by comparing model variable frequencies and odds ratios from logistic regression models in the development and validation samples (January 1, 2010-December 31, 2012, **Dataset 1**). We assessed the reliability of the clinical data elements by comparing rate of capture, and coefficients associated with each variable in the development and validation samples’ risk models.  Measure Score Reliability  For test-retest reliability, we use the randomly split development and validation samples (**Dataset 1**) and calculated the measure for each hospital separately in each sample. 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 (ICC) (Shrout and Fleiss, 1979), and assessed the values according to conventional standards (Landis and Koch, 1977). Specifically, we used **Dataset 1** split sample and calculated the risk-standardized readmission rate (RSRR) for each hospital for each sample. The agreement of the two RSRRs was quantified for hospitals using the intra-class correlation as defined by ICC (2,1) by Shrout and Fleiss (1979).  Using two independent samples provides a stringent 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 use this to estimate the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort.  Because reliability of the measure result could only be tested in a small sample of hospitals (n=21) and admissions in **Dataset 1,** testing will be repeated in a larger nationally representative set of hospitals prior to implementation. This testing will depend on implementation of hospital reporting of the core clinical data elements used in the measure’s risk models.  Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.  Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.  Rousson V, Gasser T, Seifert B. Assessing intra rater, interrater and test–retest reliability of continuous measurements. Statistics in Medicine 2002; 21:3431-3446.  Shrout P, Fleiss J. Intra class correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86:420-428.  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*)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data Element Reliability Results  For the claims-derived data elements (**Dataset 1**), the frequency of model variables remained relatively constant between 2010 and 2012, with no model variables increasing or decreasing by more than 2%.  Clinical Data Element Capture  Table 1 shows the rate of capture of all of the clinical data elements used in the measures’ risk models in the development, validation, and 2012 samples by cohort. Coefficients for the claims and clinical risk variables for all three samples can be found in data field S.2b (Data Dictionary or Code Table).  Table 1 Rate of Capture of all of the Clinical Data Elements (Dataset 1)   |  | Development Sample | Validation Sample | 2012 Sample | | --- | --- | --- | --- | | Heart rate (% captured) | | | | | Surgery/Gynecology | 95.0 | 95.2 | 96.6 | | Cardiorespiratory | 98.7 | 98.4 | 99.1 | | Cardiovascular | 97.7 | 97.9 | 98.5 | | Neurology | 97.7 | 98.1 | 98.6 | | Medicine | 98.1 | 98.1 | 98.7 | | Systolic BP (% captured) | | | | | Surgery/Gynecology | 94.5 | 94.6 | 96.0 | | Cardiorespiratory | 98.5 | 98.1 | 98.8 | | Cardiovascular | 97.6 | 97.8 | 97.9 | | Neurology | 97.7 | 98.1 | 98.5 | | Medicine | 97.9 | 97.9 | 98.5 | | Respiratory Rate (% captured) | | | | | Surgery/Gynecology | 94.4 | 94.4 | 96.1 | | Cardiorespiratory | 97.8 | 97.7 | 98.1 | | Cardiovascular | 96.8 | 97.3 | 97.3 | | Neurology | 97.0 | 97.3 | 97.6 | | Medicine | 97.1 | 97.2 | 97.6 | | Temperature (% captured) | | | | | Surgery/Gynecology | 93.7 | 94.0 | 95.7 | | Cardiorespiratory | 95.0 | 94.5 | 95.2 | | Cardiovascular | 93.6 | 93.8 | 94.3 | | Neurology | 93.1 | 94.0 | 94.5 | | Medicine | 95.1 | 95.0 | 96.0 | | Weight (% captured) | | | | | Surgery/Gynecology | 94.1 | 94.1 | 95.7 | | Cardiorespiratory | 93.7 | 93.6 | 94.9 | | Cardiovascular | 94.3 | 94.7 | 95.2 | | Neurology | 91.0 | 91.6 | 92.4 | | Medicine | 91.1 | 91.2 | 92.3 | | Oxygen Saturation (% captured) | | | | | Surgery/Gynecology | 93.3 | 93.5 | 95.8 | | Cardiorespiratory | 97.6 | 97.3 | 98.4 | | Cardiovascular | 96.1 | 96.3 | 97.4 | | Neurology | 96.2 | 96.6 | 97.4 | | Medicine | 96.0 | 95.9 | 97.3 | | Hematocrit (% captured) | | | | | Surgery/Gynecology | 83.3 | 83.8 | 82.0 | | Cardiorespiratory | 98.5 | 98.5 | 99.0 | | Cardiovascular | 95.4 | 95.5 | 94.9 | | Neurology | 97.8 | 97.9 | 98.0 | | Medicine | 97.6 | 97.6 | 98.0 | | WBC Count (% captured) | | | | | Surgery/Gynecology | 79.4 | 80.1 | 78.6 | | Cardiorespiratory | 98.5 | 98.4 | 98.9 | | Cardiovascular | 95.3 | 95.3 | 94.9 | | Neurology | 97.8 | 97.8 | 97.9 | | Medicine | 97.4 | 97.4 | 97.8 | | Potassium (% captured) | | | | | Surgery/Gynecology | 70.6 | 71.1 | 70.0 | | Cardiorespiratory | 96.8 | 96.5 | 97.1 | | Cardiovascular | 93.6 | 93.6 | 93.5 | | Neurology | 96.1 | 95.9 | 95.8 | | Medicine | 95.6 | 95.6 | 95.8 | | Sodium (% captured) | | | | | Surgery/Gynecology | 71.8 | 72.3 | 71.1 | | Cardiorespiratory | 98.7 | 98.5 | 99.1 | | Cardiovascular | 95.0 | 95.2 | 94.8 | | Neurology | 98.0 | 98.0 | 98.3 | | Medicine | 97.4 | 97.4 | 97.9 | | Bicarbonate (% captured) | | | | | Surgery/Gynecology | 71.3 | 71.7 | 70.8 | | Cardiorespiratory | 98.8 | 98.5 | 99.1 | | Cardiovascular | 95.0 | 95.3 | 94.8 | | Neurology | 98.0 | 97.9 | 98.2 | | Medicine | 97.4 | 97.4 | 97.8 | | Creatinine (% captured) | | | | | Surgery/Gynecology | 72.0 | 72.2 | 71.5 | | Cardiorespiratory | 98.7 | 98.5 | 99.1 | | Cardiovascular | 95.2 | 95.3 | 94.8 | | Neurology | 98.1 | 98.0 | 98.3 | | Medicine | 97.4 | 97.4 | 97.9 | | Glucose (% captured) | | | | | Surgery/Gynecology | 71.1 | 71.4 | 70.5 | | Cardiorespiratory | 98.6 | 98.4 | 99.0 | | Cardiovascular | 94.9 | 95.1 | 94.6 | | Neurology | 98.0 | 97.9 | 98.2 | | Medicine | 97.3 | 97.3 | 97.8 | |

Measure Score Reliability Results

In **Dataset 1**, there were 81,589 in the development sample and 79,813 in the validation sample. The agreement between the two RSRRs for each hospital was 0.688, which according to the conventional interpretation is “moderate” (Landis & Koch, 1977).

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

|  |
| --- |
| The stability over time of the odds ratios or variable coefficients and the model variable frequencies and rates of capture for clinical data elements suggests that the underlying data elements are reliable.  Measure Testing  For the hospital event rate based on the patient binomial outcomes like readmission (Yes/No), an ICC value of 0-0.2 indicates poor agreement; 0.3-0.4 indicates fair agreement; 0.5-0.6 indicates moderate agreement; 0.7-0.8 indicates strong agreement; and >0.8 indicates almost perfect agreement. The ICC of 0.688 is moderate to strong in **Dataset 1.** |

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

**2b2. VALIDITY TESTING**

**2b2.1. 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*)

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

|  |
| --- |
| Validity of EHR Data Elements  Several critical clinical data elements used in the measure’s risk models were derived from patients’ electronic medical records. When this measure is implemented, CMS intends to obtain these critical data elements from hospital EHRs and merge the data with claims data to calculate and report measure results. We tested the validity of electronic extraction of these critical data elements as part of a more comprehensive evaluation of a larger set of core clinical data elements (CCDEs). The CCDE are a set of 21 EHR data elements that are captured on most adults (plus Troponin, which is a condition-specific CCDE for patients with acute myocardial infarction) admitted to acute care hospitals, are easily extracted from EHRs, and can be used to risk adjust hospital outcome measures for a variety of conditions and procedures. All of the critical data elements used in the Hybrid HWR measure are included in the CCDE. Testing of the CCDE involved three phases: 1) identification of potentially feasible clinical data through qualitative assessment, 2) empirical feasibility testing of several clinical data elements electronically extracted from two large multi-facility health systems, and 3) validity testing of the CCDE at an additional health system.  **Phase 1: Identification of potentially feasible clinical data through qualitative assessment**  In order to identify the CCDEs for risk adjustment of hospital outcome measures for adult patients, we first conducted a qualitative assessment of the reliable capture, accuracy, and extractability of categories and subcategories of clinical data as defined by the Quality Data Model (QDM) (e.g., vital signs, laboratory test results). We established a set of criteria to assess the consistency of data capture, relevance to hospital quality measures, and extractability from health records.  Data Capture Criteria:   * Obtained consistently under current practice. Routinely collected for patients admitted to the hospital under current clinical practice and EHR workflows. * Captured with a standard definition. Consistent conceptual understanding, method of collection, and units of measurement. * Entered in a structured field. Captured in numerical, pseudo-numerical, or list format. * Data Extraction Criteria: * Encoded consistently. Can be linked to a standard and uniform coding structure such as ICD-9 or Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT). * Extractable from the EHR. Can be readily and consistently identified and exported from current EHR databases. * Exported with metadata. Additional information such as time stamps and reference values that are needed for interpretation are consistently available.   These criteria are aligned with those established in the NQF’s eMeasure Feasibility Assessment Report as well as the NQF feasibility criteria (see included Data Element Feasibility Scorecard). The NQF report emphasized four key aspects of feasibility. First, data should be structured or easily converted to a structured and interpretable format. Second, data should be accurate. Third, data should be easily associated with a standard set of codes to ensure consistent extraction across EHR environments. Finally, data should not require changes to current clinical practice or workflows.  We then convened a Technical Expert Panel (TEP) to apply these criteria to categories and subcategories (data types) of clinical data based on the Quality Data Model (QDM). We asked TEP members to consider only the context of adult hospitalized patients when making their assessments. Data categories and subcategories were rated on each feasibility criterion independently by TEP members. The ratings were tallied and TEP members met to discuss and resolve areas of disagreement. Through this process the TEP identified a list of data subcategories that were potentially feasible for use in hospital outcome measures. The CCDE were derived from only those subcategories for which the TEP reached consensus agreement on feasibility.  **Phase 2: Empirical feasibility testing using a large multi-site database**  In **Dataset 1**, we next directly examined the feasibility of clinical data elements from the subcategories identified by the TEP as feasible (for all adult inpatient admissions). We examined all admissions in **Dataset 1** between 2010 and 2011. We analyzed clinical data elements to determine whether they were captured in a numerical field, the consistency and timing of capture, and the accuracy of the data elements. We examined the data elements across conditions, hospitals, and point of hospital entry. We tested several data elements that met the feasibility criteria in models predicting 30-day mortality following admission for several common medical conditions. The complete list of 21 (plus Troponin) CCDE were derived from these analyses.  To verify that the findings from our analysis of **Dataset 1** were generalizable to other hospitals and electronic health systems, we partnered with Premier Inc., a collaborative healthcare alliance of approximately 2,900 U.S. community hospitals focused on measuring and improving their members’ quality outcomes and safely reducing healthcare costs. We administered a survey to four of their member hospital systems that used a variety of EHR systems to confirm the availability of the clinical data elements. Additionally, we assessed the rate and timing of capture of the data elements identified in **Dataset 1** in ERH abstracted from a second health system consisting of three hospitals (**Dataset 2**).  **Phase 3: Validity testing at two hospital sites the CCDE (including critical data elements for the Hybrid HWR measure)**  In Phase 3, we developed electronic specifications (e-specifications) using the Measure Authoring Tool (MAT), and analyzed extracted data from EHRs. We assessed the ability of hospitals to use the e-specifications to query and electronically extract CCDEs from the EHR, for all adult inpatient admissions occurring over the course of one year. Validity testing assessed the accuracy of the electronically extracted CCDEs compared to the same CCDEs gathered through manual abstraction (from the EHR) in a subset of 368 charts identified in the data query in **Dataset 2**, and 391 charts identified in the data query in **Dataset 3**.  *Chart Abstraction*: We calculated the number of admissions that needed to be randomly sampled from the EHR dataset and manually abstracted to yield a statistical margin of error (MOE) of 5% and a confidence level of 95% for the match rates between the two data sources. Sites then used an Access-based manual abstraction tool provided (along with training) to manually abstract the CCDEs from the random samples of the medical records identified through the EHR data query. The manual chart abstraction data is considered the “gold standard” for the purpose of this analysis.  *Validity Testing*: We conducted validity testing on the critical EHR data elements in the Hybrid HWR measure. For each continuous data element, we were only interested in the case where the electronic abstraction value exactly matched the manual abstraction value. We therefore only calculated the raw agreement rate between data from electronic and manual chart abstraction. For simple data values, we believe taking this approach, as compared to reporting statistical tests of accuracy, better reflects the concept of matching exact data values rather than calculated measure results. Therefore, we do not report statistical testing of the accuracy of the EHR derived data value as compared with the abstracted value. Instead, we counted only exact matches in the data value as well as the time and date stamp associated with that value when we calculated the match rate. The 95% confidence level was established based on the sample size and reflects the exact match rate using these criteria.  Validation Against Other Risk Models and Registry Data  The Hybrid model we developed uses a combination of claims data (demographics, comorbidities, and patient medical history) and electronic clinical data (laboratory results and vital signs).  We compared the Hybrid risk model to the harmonized claims-only risk model used in the publicly reported Hospital-Wide All Cause Unplanned Readmission Measure. Both models use inpatient administrative claims data to derive the cohort, to derive risk variables, and to assess the unplanned readmission outcome.  Measure validity was tested through comparison of this Hybrid risk adjustment model with claims-only risk-adjustment model, and through use of established measure development guidelines.  For the derivation of both risk models, we used **Dataset 1**. Both the Hybrid and claims-only risk models used the same inclusion/exclusion criteria and a risk-adjustment (statistical modeling) strategy and only differed with respect to the risk variables used. We compared the model discrimination and the correlation in hospital performance results for the two models.  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  Following completion of the preliminary model, we solicited public comment on the measure through the CMS site link: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/CallforPublicComment.html>. 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.  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  Suter L, Wang C, Araas M, et al. Hospital-Level 30-Day All-Cause Unplanned Readmission Following Coronary Artery Bypass Graft Surgery (CABG): Updated Measure Methodology Report. 2014; http://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228890352615&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DRdmsn\_CABG\_MeasMethd\_Rpt\_060314.pdf&blobcol=urldata&blobtable=MungoBlobs. Accessed November 4, 2015.  ICD-9 to ICD-10 Conversion  Statement of Intent  [X] Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.  [ ] Goal was to take advantage of the more specific code set to form a new version of the measure, but fully consistent with the original intent.  [ ] The intent of the measure has changed.  Process of Conversion  ICD-10 codes were initially 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 currently in use for this measure. An ICD-9 to ICD-10 crosswalk is attached in field S.2b. (Data Dictionary or Code Table).  We have also examined the updated ICD-9 Map to AHRQ Conditions Categories (CCS) crosswalk to the ICD-10 CCS map provided by AHRQ in preparation for the inclusion of ICD 10 data in this measure. Please refer to the ICD-10 CCS map on the [AHRQ](https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp) website. |

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

|  |
| --- |
| Validity of EHR Data Elements  **Phase 1: TEP Survey Results**  The TEP identified seven subcategories of EHR data that they considered feasible for adult hospitalized patients. They were: Encounter Performed, Patient Characteristics including birth date and sex, Physical Examination Findings for vital signs only, Diagnostic Study Order, Diagnostic Study Performed, Medication Discharge, and Laboratory Test Result. We limited the CCDE to data elements to only four categories: Encounter Performed, Patient Characteristics, Physical Examination Findings for vital signs only, and Laboratory Test Results, which are unlikely to be reflective of care quality and therefore are thought to be both feasible to extract and appropriate for risk adjustment.  **Phase 2: Feasibility Testing Results**  **Datasets 1 and 2**: The consistency of data capture of the critical data elements included in the Hybrid HWR measure for all adult hospitalized patients in **two health systems** with different EHR environments (EPIC and Cerner). Results presented in Table 2 and Table 3 are from EPIC system; results presented in Table 4 are from Cerner system. These tables show consistent capture of all the clinical data elements used in the risk models.  In addition, the four Premier member hospitals all reported that the CCDE were: captured in the inpatient EHR; captured in the emergency department EHR; recorded in a structured format; extracted for reporting; extracted for other purposes; and time and date stamps capture. |

Table 2. Proportion of Episodes with Captured Vital Signs at Various Time points (Dataset 1)

| Vital Sign Finding – Full Cohort | Total with Finding and Timestamp  % | Within 2 Hours  % | Within 6 Hours  % | Within 12 Hours  % |
| --- | --- | --- | --- | --- |
| Basic vital signs | | | | |
| Heart rate | 99.7 | 96.8 | 99.4 | 99.6 |
| Systolic blood pressure | 99.7 | 96.7 | 99.3 | 99.6 |
| Diastolic blood pressure | 99.7 | 96.7 | 99.3 | 99.6 |
| Respiratory rate | 99.7 | 95.8 | 99.1 | 99.6 |
| Temperature | 99.7 | 93.7 | 98.5 | 99.5 |
| Oxygen saturation | 98.2 | 86.0 | 92.6 | 95.4 |
| Weight | 92.5 | 80.2 | 85.2 | 88.8 |

Table 3. Proportion of Admissions with Laboratory Results at Various Time Points (Dataset 1)

| Lab Test Result – Full Cohort | Total with Result and Timestamp (%) | | Within 2 Hours (%) | Within 6 Hours (%) | Within 12 Hours (%) | Within 24 Hours (%) |
| --- | --- | --- | --- | --- | --- | --- |
| Hemoglobin | 92.7 | 61.2 | | 72.7 | 77.3 | 90.6 |
| Hematocrit | 92.8 | 61.6 | | 73.8 | 78.0 | 90.8 |
| Platelets | 92.0 | 61.1 | | 72.4 | 76.5 | 89.8 |
| WBC count | 92.0 | 61.1 | | 72.4 | 76.5 | 89.8 |
| Potassium | 71.3 | 49.3 | | 57.2 | 60.2 | 69.4 |
| Sodium | 71.6 | 49.3 | | 57.3 | 60.3 | 69.6 |
| Chloride | 71.1 | 49.3 | | 56.1 | 59.4 | 69.0 |
| Bicarbonate | 71.2 | 49.2 | | 56.8 | 59.8 | 69.2 |
| Glucose | 72.0 | 49.7 | | 57.6 | 60.6 | 70.0 |
| Troponin | 32.2 | 25.6 | | 28.7 | 30.0 | 30.6 |

Table 4. Percent Captured per Data Element per Hospital (Dataset 2)

| Data Element/ CCDE | % Captured Hospital 1 | % Captured Hospital 2 | % Captured Hospital 3 |
| --- | --- | --- | --- |
| Heart Rate (BPM) | 94.27 | 80.54 | 84.48 |
| Systolic Blood Pressure (mmHG) | 94.42 | 80.88 | 84.07 |
| Diastolic Blood Pressure (mmHG) | 94.29 | 80.81 | 83.99 |
| Respiratory Rate (BPM) | 93.58 | 87.69 | 84.26 |
| Temperature (C) | 92.43 | 86.48 | 83.77 |
| Oxygen Saturation (%) | 83.22 | 78.80 | 83.61 |
| Weight (Kg) | 99.01 | 98.66 | 98.61 |
| Hemoglobin | 96.19 | 96.29 | 91.47 |
| Hematocrit | 95.77 | 96.29 | 91.25 |
| Platelet | 89.40 | 96.52 | 92.26 |
| WBC Count | 89.12 | 96.25 | 91.11 |
| Potassium | 78.44 | 81.08 | 82.03 |
| Sodium | 78.29 | 81.03 | 81.78 |
| Chloride | 78.28 | 81.03 | 81.78 |
| Bicarbonate | 11.53 | 11.23 | 12.29 |
| Glucose | 77.64 | 80.93 | 81.32 |
| Troponin | 29.67 | 32.10 | 15.71 |

**Phase 3: Further Feasibility and Validity Testing Results**

Chart abstraction for validity testing was done in **Dataset 2** and **Dataset 3.** Table 5 demonstrates the comparison between electronic and manual abstraction of data in the two health systems.

Table 5. Percent Agreement and Confidence Internals (Dataset 2 and 3)

| Data Element/ CCDE | % Agreement Between Datasets (Number Matching/ Total Records With A Data Value) | | 95% Confidence Internal for Agreement | % Present in Electronic Extraction, Missing in Manual Abstraction (N) | % Present in Manual Abstraction, Missing in Electronic Extraction (N) | % Missing in Both Electronic Extraction and Manual Abstraction (N) |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset 2 (n=368) | | | | | | |
| Heart rate (BPM) | 95.55 (322/337) | 92.76 - 97.49 | | 0 (0.00) | 8.42 (31) | 0 (0.00) |
| Syst Blood Pressure (mmHG) | 94.67 (320/338) | 91.71 - 96.81 | | 0 (0.00) | 8.15 (30) | 0 (0.00) |
| Diast Blood Pressure (mmHG) | 94.38 (319/338) | 91.36 - 96.58 | | 0 (0.00) | 8.15 (30) | 0 (0.00) |
| Respiratory Rate (BPM) | 94.89 (316/333) | 91.95 - 97.00 | | 0 (0.00) | 9.51 (35) | 0 (0.00) |
| Temperature (C) | 95.41 (312/327) | 92.55 - 97.41 | | 0 (0.00) | 10.87 (40) | 0.27 (1) |
| Oxygen Saturation (%) | 94.68 (285/301) | 91.51 - 96.93 | | 0.27 (1) | 14.40 (53) | 3.53 (13) |
| Weight (Kg)\* | 14.66 (51/348) | 11.11 - 18.81 | | 1.09 (4) | 3.53 (13) | 0.82 (3) |
| Hemoglobin (g/dL) | 96.50 (331/343) | 93.97 - 98.18 | | 0.82 (3) | 3.80 (14) | 2.17 (8) |
| Hematocrit (%) | 96.19 (328/341) | 93.57 - 97.95 | | 0.82 (3) | 3.80 (14) | 2.72 (10) |
| Platelet (x10(9)/L) | 96.88 (310/320) | 94.33 - 98.49 | | 0.82 (3) | 4.62 (17) | 7.61 (28) |
| WBC Count (x10(9)/L) | 96.56 (309/320) | 93.93 - 98.27 | | 0.82 (3) | 4.62 (17) | 7.61 (28) |
| Potassium (meq/L) | 97.22 (280/288) | 94.60 - 98.79 | | 0 (0.00) | 5.16 (19) | 16.58 (61) |
| Sodium (meq/L) | 97.21 (279/287) | 94.58 - 98.79 | | 0 (0.00) | 5.43 (20) | 16.58 (61) |
| Chloride (meq/L) | 97.21 (279/287) | 94.58 - 98.79 | | 0 (0.00) | 5.43 (20) | 16.58 (61) |
| Bicarbonate (meq/L) | 14.81 (8/54) | 6.62 - 27.12 | | 0.27 (1) | 68.21 (251) | 16.85 (62) |
| Glucose (mg/dL) | 96.14 (274/285) | 93.20 - 98.06 | | 0 (0.00) | 5.43 (20) | 17.12 (63) |
| Troponin (ng/mL) | 93.33 (98/105) | 86.75 - 97.28 | | 4.08 (15) | 0.54 (2) | 66.85 (246) |
| Dataset 3 (n=391) | | | | | | |
| Heart rate (BPM) | 57.45 (135/235) | 50.85 - 63.85 | | 0 (0.00) | 39.39 (154) | 0.51 (2) |
| Syst Blood Pressure (mmHG) | 60.26 (138/235) | 53.61 - 66.65 | | 0 (0.00) | 40.92 (160) | 0.51 (2) |
| Diast Blood Pressure (mmHG) | 60.09 (137/228) | 53.41 - 66.50 | | 0.26 (1) | 40.92 (160) | 0.51 (2) |
| Respiratory Rate (BPM) | 70.14 (155/221) | 63.63 - 76.09 | | 0 (0.00) | 42.71 (167) | 0.77 (3) |
| Temperature (C) | 79.09 (174/220) | 73.11 - 84.27 | | 0 (0.00) | 42.46 (166) | 1.28 (5) |
| Oxygen Saturation (%) | 56.65 (115/203) | 49.53 - 63.57 | | 0.26 (1) | 46.80 (183) | 1.02 (4) |
| Weight (Kg) | 84.41 (157/186) | 78.38 - 89.30 | | 0.26 (1) | 48.34 (189) | 3.84 (15) |
| Hemoglobin (g/dL) | 88.78 (87/98) | 80.80 - 94.26 | | 0.26 (1) | 58.06 (227) | 16.62 (65) |
| Hematocrit (%) | 91.67 (88/96) | 84.24 - 96.33 | | 0.26 (1) | 58.31 (228) | 16.88 (66) |
| Platelet (x10(9)/L) | 94.68 (89/94) | 88.02 - 98.25 | | 0.26 (1) | 59.08 (231) | 16.62 (65) |
| WBC Count (x10(9)/L) | 94.62 (88/93) | 87.90 - 98.23 | | 0.26 (1) | 59.34 (232) | 16.62 (65) |
| Potassium (meq/L) | 86.75 (72/83) | 77.52 - 93.19 | | 0 (0.00) | 62.66 (245) | 16.11 (63) |
| Sodium (meq/L) | 91.46 (75/82) | 83.20 - 96.50 | | 0 (0.00) | 61.64 (241) | 17.39 (68) |
| Chloride (meq/L) | 97.56 (80/82) | 91.47 - 99.70 | | 0 (0.00) | 60.61 (237) | 18.41 (72) |
| Bicarbonate (meq/L) | 29.41 (5/17) | 10.31 - 55.96 | | 0.26 (1) | 77.49 (303) | 17.90 (70) |
| Glucose (mg/dL) | 95.12 (78/82) | 87.98 - 98.66 | | 0 (0.00) | 63.17 (247) | 15.86 (62) |
| Troponin (ng/mL) | 82.61 (19/23) | 61.22 - 95.05 | | 0 (0.00) | 13.30 (52) | 80.82 (316) |

|  |
| --- |
| A post-validation review of the code used by the hospital in **Dataset 3**, revealed that the hospital experienced a number of errors. The most significant of which was extracting data only within an incorrect two-hour window for laboratory test results (the correct window was 24 hours). Additionally, physical exam (vital signs) data were extracted based on the date/time that results were documented rather than the date/time the physical exams were performed, driving down the accuracy of these data. However, post-validation review of the code used by the hospital in **Dataset 2** showed no such errors in the query executed. As a result the match rate was much higher.  Validation against Claims-Only Risk Model **(Dataset 1)**  We estimated hospital-level RSRRs using the corresponding hierarchical logistic regression for each of the models in the linked patient sample. We then examined the linear relationship between the estimates using regression techniques and weighting by the total number of cases in each hospital. The Pearson correlation coefficient of the standardized rates from the claims-only risk-adjustment model and the Hybrid risk-adjustment model in the Development Sample of **Dataset 1** is 0.9902. |

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

|  |
| --- |
| Validity of EHR Data Elements  Feasibility Testing (Phases 1-3)  The critical data elements were demonstrated to be feasible through consensus of the TEP and direct examination of EHR data establishing consistent capture of the CCDE among adult hospitalized patients. In addition, we established the validity of electronic extraction of the CCDE demonstrated by the high match rate when comparing EHR extracted and manual medical record abstracted CCDE values.  Measure Validity (**Dataset 1**)  The results between the Hybrid HWR model and the claims-only risk model were nearly identical. In addition, the high correlation among the RSRRs calculated from all models shows that each model provides a similar or consistent measure result for hospitals. |

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

**2b3. EXCLUSIONS ANALYSIS**

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

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

|  |
| --- |
| All exclusions were determined by careful clinical review and have been made based on clinically relevant decisions and 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 (**Dataset 1**). These exclusions are consistent with similar NQF-endorsed outcome measures. Rationales for the exclusions are detailed in data field S.10 (Denominator Exclusions). |

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| For the purposes of tabulation, exclusions are performed sequentially. Thus, a hospital stay that would be excluded based on multiple criteria is counted in the first criterion only. This data is from the original initial cohort (n= 251,006) of index admissions.   |  |  |  | | --- | --- | --- | | **Exclusion** | **N** | **%** | | Discharged Against Medical Advice | 679 | 0.27% | | Cancer Treatment | 6,356 | 2.53% | | Psychiatric Treatment | 593 | 0.24% | | Rehabilitation | 885 | 0.35% |   These exclusions represent only 3.88% of the initial cohort. We do not report frequency of distribution of exclusions across measured entities due to the minimal impact of the exclusions on the measure cohort. The final cohort was 242,515 index admissions. |

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

|  |
| --- |
| 1. Patients discharged against medical advice (AMA) account for 0.27% 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 and prepare the patient for discharge. 2. Patients admitted for primary psychiatric diagnoses account for 0.24% of all index admissions excluded from the initial cohort. This exclusion is needed because these patients are typically cared for in separate psychiatric or rehabilitation centers which are not comparable to acute care hospitals. 3. Patients admitted for rehabilitation account for 0.35% of all index admissions excluded from the initial cohort. This exclusion is needed because patients admitted for rehabilitation are not admitted for treatment of acute illness and the care provided in rehabilitation centers is not comparable to care provided in acute care hospitals. 4. Patients admitted for medical treatment of cancer account for 2.53% of all index admissions excluded from the initial cohort. Admissions for treatment of cancer are associated with a very different mortality and readmission risk compared with admissions to other Inpatient Prospective Payment Systems (IPPS) hospitals for treatment of other diseases. Additionally, outcomes for these admissions do not correlate well with outcomes for other types of admissions. (Patients with cancer who are admitted for other diagnoses or for surgical treatment of their cancer remain in the measure). |

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

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

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

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

**2b4.2. If an outcome or resource use 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 |

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient 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 and not related to disparities*)

|  |
| --- |
| Our approach to risk adjustment was tailored to and appropriate for a publicly reported outcome measure, as articulated in the American Heart Association (AHA) Scientific Statement, “Standards for Statistical Models Used for Public Reporting of Health Outcomes” (Krumholz et al., 2006).  The measure estimates hospital-level 30-day all-cause RSRRs using hierarchical logistic regression models. In brief, the approach simultaneously models data at the patient and hospital levels to account for variance in patient outcomes within and between hospitals (Normand S-LT, Shahian DM, 2007). At the patient level, it models the log-odds of hospital readmission within 30 days of discharge using age, selected clinical covariates, and a hospital-specific intercept. At the hospital level, the approach models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept represents the underlying risk of a readmission at the hospital, after accounting for patient risk. The hospital-specific intercepts are given a distribution to account for the clustering (non-independence) of patients within the same hospital (Normand S-LT, Shahian DM., 2007). If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.  Admissions are assigned to one of five mutually exclusive specialty cohort groups consisting of related conditions or procedures. For each specialty cohort group, the standardized readmission ratio (SRR) is calculated as the ratio of the number of “predicted” readmissions to the number of “expected” readmissions at a given hospital. For each hospital, the numerator of the ratio is the number of readmissions within 30 days predicted based on the hospital’s performance with its observed case mix and service mix, and the denominator is the number of readmissions expected based on the nation’s performance with that hospital’s case mix and service mix. This approach is analogous to a ratio of “observed” to “expected” used in other types of statistical analyses. It conceptually allows a particular hospital’s performance, given its case mix and service mix, to be compared to an average hospital’s performance with the same case mix and service mix. Thus, a lower ratio indicates lower-than-expected readmission rates or better quality, while a higher ratio indicates higher-than-expected readmission rates or worse quality.  For each specialty cohort, the “predicted” number of readmissions (the numerator) is calculated by using the coefficients estimated by regressing the risk factors (found in the attached Data Dictionary) and the hospital-specific intercept on the risk of readmission. The estimated hospital-specific intercept for each cohort is added to the sum of the estimated regression coefficients multiplied by patient characteristics. The results are transformed and summed over all patients attributed to a hospital to get a predicted value. The “expected” number of readmissions (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 transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimate the model coefficients using the data in that period.  The specialty cohort SRRs are then pooled for each hospital using a volume-weighted geometric mean to create a hospital-wide composite SRR. The composite SRR is multiplied by the national observed readmission rate to produce the RSRR.  Data Source  To select candidate variables for the Hybrid risk model, we began with the list of all administrative claims-based risk-adjustment variables included in the currently publicly reported Hospital-Wide All Cause Unplanned Readmission Measure. These candidate variables were derived from: the index admission, with comorbidities identified from the index admission secondary diagnoses (excluding potential complications) and 12-month pre-index inpatient data (for any condition). In identifying these variables for the current publically reported HWR measure, we sought to develop a model that was parsimonious, using a grouper that is in the public domain for the 15,000+ ICD-9-CM codes we started with the 189 diagnostic groups included in the Hierarchical Condition Category (HCC) clinical classification system (Pope et al., 2000). The HCC clinical classification system was developed for CMS in preparation for all-encounter risk adjustment for Medicare Advantage (managed care) plans and represented a refinement of an earlier risk-adjustment method based solely on principal inpatient diagnosis. The HCC model makes use of all physician and hospital encounter diagnoses and was designed to predict a beneficiary’s expenditures based on the total clinical profile represented by all of his/her assigned HCCs. Under the HCC algorithm, the 15,000+ ICD-9-CM diagnosis codes are first assigned to one of 804 mutually exclusive groupings (“DxGroups”) and then subsequently aggregated into 189 condition categories (CCs). During development, we used the April 2008 version of the ICD-9-CM to CC assignment map, which is maintained annually by CMS and posted at www.qualitynet.org. We do not use the hierarchy and therefore refer to the CCs rather than HCCs. The HWR risk-adjustment models use only inpatient claims data (history and current) in order to make it feasible to implement with Medicare data, and to make it applicable to all-payer data, which are typically restricted to inpatient claims.  We also used the core clinical data elements CCDE, the EHR-derived data elements used in the measure. The CCDE include the first vital signs captured within 2 hours of the start of the encounter and the first set of results for several basic laboratory tests captured within 24 hours of the start of the encounter (for example, complete blood count and basic chemistry panel). A file that contains a list of the ICD-9-CM codes and their groupings into CCs as well as a list of the CCDE is attached in data field S.2b (Data Dictionary and Code Table).  Complications occurring during hospitalization are not comorbid illnesses, may reflect hospital quality of care, and therefore should not be used for risk adjustment. Hence, conditions that may represent adverse outcomes due to care received during the index hospital stay are not included in the risk-adjusted model.  Service mix adjustment: The measure includes many different discharge condition categories that differ in their baseline readmission risks. In addition, hospitals differ in their relative distribution of these condition categories (service mix). To adjust for service mix, the measure uses an indicator variable for the discharge condition category in addition to risk variables for comorbid conditions. The models include a condition-specific indicator for all condition categories with sufficient volume (defined as those with more than 1,000 admissions nationally in a given year for Medicare FFS data) as well as a single indicator for conditions with insufficient volume in each model.  Although the 5 risk models use a common set of claims variables, the CCDE variables are not the same across specialty cohort models. Only those data elements that are statistically significant in each individual model are included. We estimate a hierarchical logistic regression model for each specialty cohort separately, and the coefficients associated with each variable may vary across specialty cohorts.  The final set of risk-adjustment variables is listed on the Submission form in item S.14 and in the Data Dictionary and Code Table attached in data filed S.2b  References:  Krumholz HM, Brindis RG, Brush JE, et al. 2006. 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 113: 456-462.  Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.  Pope, G., et al., Principal Inpatient Diagnostic Cost Group Models for Medicare Risk Adjustment. Health Care Financing Review, 2000. 21(3):26.  Socioeconomic Status (SES) Factors and Race  We selected variables representing SES factors and race for examination based on a review of literature, conceptual pathways, and feasibility. In Section 1.8, we describe the variables that we considered and analyzed based on this review. Below we describe the pathways by which SES and race may influence 30-day readmission.  Our conceptualization of the pathways by which patient SES or race affects 30-day readmission is informed by the literature.  Literature Review of Socioeconomic Status (SES) and Race Variables and HF Readmission  To examine the relationship between SES and race variables and hospital 30-day, hospital-wide, all-cause, unplanned readmission following hospitalization, a literature search was performed with the following exclusion criteria: international studies, articles published more than 10 years ago, articles without primary data, articles using Veterans Affairs databases as the primary data source, and articles not explicitly focused on SES or race and readmission across multiple conditions. One hundred and sixty nine articles were initially reviewed, and 155 studies were excluded from full-text review based on the above criteria. Studies indicate that SES and race variables were associated with increased risk of readmission across multiple major illnesses and conditions (Aseltine RH, et al., 2015; Mitchell SE, et al., 2012; Odonkor CA, et al., 2015; Herrin J, et al., 2015; Gu Q, et al., 2014, Kim H, et al., 2010; Kangovi S, et al., 2012; Iloabuchi TC, 2014; Beck AF, et al., 2012; Arbaje AI, et al., 2008; Hu J, 2014; Nagasako EM, et al., 2014; Joynt, K. E., et al., 2013), though there may not be a significant effect on hospital-level profiling (Blum et al., 2014).  Causal Pathways for Socioeconomic Status (SES) and Race Variable Selection  Although some recent literature evaluates the relationship between patient SES or race and the readmission outcome, few studies directly address causal pathways or examine the role of the hospital in these pathways. 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 readmission. The SES factors that have been examined in the readmission 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 range from the self-reported or documented race or ethnicity of the patient to the patient’s income or education level (Eapen et al., 2015; Hu et al., 2014). Neighborhood/community-level variables use information from sources such as the American Community Survey (ACS) 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). 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 and Jha, 2013).  The conceptual relationship, or potential causal pathways by which these possible SES risk factors influence the risk of readmission 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. **Relationship of socioeconomic status (SES) factors or race to health at 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 SES 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 childcare), 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**.**  In addition to SES risk factors, studies have shown that worse health status is more prevalent among African-American patients compared with white patients. The association between race and worse health is in part mediated by the association between race and SES risk factors such as poverty or disparate access to care associated with poverty or neighborhood. The association is also mediated through bias in healthcare as well as other facets of society.  2. **Use of low-quality hospitals**. Patients of lower income, lower education, or unstable housing have been shown not to have equitable access to high quality facilities 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 contribute to increased risk of readmission following hospitalization (Jha et al., 2011; Reames et al., 2014). Similarly African-American patients have been shown to have less access to high quality facilities compared with white patients (Skinner et al., 2005).  3. **Differential care within a hospital**. The third major pathway by which SES factors or race may contribute to readmission risk is that patients may not receive equivalent care within a facility. For example, African-American patients have been shown to experience differential, lower quality, or discriminatory care within a given facility (Trivedi et al., 2014). Alternatively, patients with SES risk factors such as lower education may require differentiated care – e.g. provision of lower literacy information – that they do not receive.  4. **Influence of SES on readmission risk outside of hospital quality and health status**. Some SES risk factors, such as income or wealth, may affect the likelihood of readmission 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 economic priorities or a lack of access to care outside of the hospital.  These proposed pathways are complex to distinguish analytically. They also have different implications on the decision to risk adjust or not. We, therefore, first assessed if there was evidence of a meaningful effect on the risk model to warrant efforts to distinguish among these pathways. Based on this model and the considerations outlined in Section 1.8, the following SES and race variables were considered:  • Dual eligible status  • African American race  • AHRQ SES index  Using the data from the Hospital-Wide All-Cause Readmission Measure for the 2015 reporting year (**Dataset 4** and **Dataset 5** [AHRQ SES index]), we assessed the relationship between the SES variables and race 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 SES or race factors 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 readmission 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 readmission rates (hospital-level effect). Thus, as an additional step, we performed a decomposition analysis to assess the independent effects of the SES and race variables at the patient level and the hospital level. If, for example, all the elevated risk of readmission for patients of low SES was due to lower quality/higher readmission 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 readmission risk was 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 decomposed each of the SES and race variables as follows: Let Xij be a binary indicator of the SES or race status of the ith patient at the jth hospital, and **X**j the percent of patients at hospital j with Xij = 1. Then we rewrote Xij = (Xij- **Xj**) + **X**j ≡ Xpatient+ Xhospital. The first variable, Xpatient, represents the effect of the risk factor at the patient level (sometimes called the “within” hospital effect), and the second, 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, 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 or African-American patients on the readmission rate of an average patient; and 2) a patient’s SES or race on their own readmission 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 or race in the model is binary whereas the hospitals’ proportion of low SES patients or African-American patients is continuous.  References:  Arbaje AI, Wolff JL, Yu Q, Powe NR, Anderson GF, Boult C. Postdischarge environmental and socioeconomic factors and the likelihood of early hospital readmission among community-dwelling Medicare beneficiaries. *The Gerontologist.* 2008;48(4):495-504.  Aseltine RH, Jr., Yan J, Gruss CB, Wagner C, Katz M. Connecticut Hospital Readmissions Related to Chest Pain and Heart Failure: Differences by Race, Ethnicity, and Payer. *Connecticut medicine.* 2015;79(2):69-76.  Beck AF, Simmons JM, Huang B, Kahn RS. Geomedicine: area-based socioeconomic measures for assessing risk of hospital reutilization among children admitted for asthma. *American journal of public health.* 2012;102(12):2308-2314.  Blum, A. B., N. N. Egorova, E. A. Sosunov, A. C. Gelijns, E. DuPree, A. J. Moskowitz, A. D. Federman, D. D. Ascheim and S. Keyhani. "Impact of Socioeconomic Status Measures on Hospital Profiling in New York City." *Circ Cardiovasc Qual Outcomes* 7, no. 3 (2014): 391-7.  Eapen ZJ, McCoy LA, Fonarow GC, Yancy CW, Miranda ML, Peterson ED, Califf RM, Hernandez AF. Utility of socioeconomic status in predicting 30-day outcomes after heart failure hospitalization. Circ Heart Fail. May 2015; 8(3):473-80.  Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. 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.  Gu Q, Koenig L, Faerberg J, Steinberg CR, Vaz C, Wheatley MP. The Medicare Hospital Readmissions Reduction Program: potential unintended consequences for hospitals serving vulnerable populations. *Health services research.* 2014;49(3):818-837.  Herrin J, St Andre J, Kenward K, Joshi MS, Audet AM, Hines SC. Community factors and hospital readmission rates. *Health services research.* 2015;50(1):20-39.  Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health affairs (Project Hope). 2014; 33(5):778-785.  Iloabuchi TC, Mi D, Tu W, Counsell SR. Risk factors for early hospital readmission in low-income elderly adults. *Journal of the American Geriatrics Society.* 2014;62(3):489-494.  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, K. E., E. J. Orav and A. K. Jha. "Thirty-Day Readmission Rates for Medicare Beneficiaries by Race and Site of Care." JAMA 305, no. 7 (2011): 675-81.  Kangovi S, Grande D, Meehan P, Mitra N, Shannon R, Long JA. Perceptions of readmitted patients on the transition from hospital to home. *Journal of hospital medicine.* 2012;7(9):709-712.  Kim H, Ross JS, Melkus GD, Zhao Z, Boockvar K. Scheduled and unscheduled hospital readmissions among patients with diabetes. *The American journal of managed care.* 2010;16(10):760-767.  Mitchell SE, Sadikova E, Jack BW, Paasche-Orlow MK. Health literacy and 30-day postdischarge hospital utilization. *Journal of health communication.* 2012;17 Suppl 3:325-338.  Nagasako EM, Reidhead M, Waterman B, Dunagan WC. Adding socioeconomic data to hospital readmissions calculations may produce more useful results. *Health affairs (Project Hope).* 2014;33(5):786-791.  Odonkor CA, Hurst PV, Kondo N, Makary MA, Pronovost PJ. Beyond the Hospital Gates: Elucidating the Interactive Association of Social Support, Depressive Symptoms, and Physical Function with 30-Day Readmissions. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists.* 2015;94(7):555-567.  Reames BN, Birkmeyer NJ, Dimick JB, Ghaferi AA. Socioeconomic disparities in mortality after cancer surgery: failure to rescue. JAMA surgery 2014; 149:475-81.  Skinner J, Chandra A, Staiger D, Lee J, McClellan M. Mortality after acute myocardial infarction in hospitals that disproportionately treat black patients. Circulation 2005; 112:2634-41.  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. |

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

|  |
| --- |
| Candidate and final model variables, with a corresponding CC to ICD-9 code map, are listed in the accompanying Excel Data Dictionary. The model variables from the original HWR measure are forced into the final model to align with that measure. The CCDE variables included in the model use logistic regression models with stepwise selection method (P=0.05). |

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

|  |
| --- |
| Variation in prevalence of the factor across measured entities  The prevalence of SDS factors in the claims-only HWR cohort varies across measured entities. The median percentage of dual eligible patients is 14.9% (interquartile range [IQR] 9.8%-22.6 %). The median percentage of African-American patients is 2.2% (IQR 0.0%-9.4%). The median percentage of low SES AHRQ indicator patients is 19.4% (IQR 5.0%-57.3%).  Empirical association with the outcome (bivariate)  The patient-level observed hospital-wide readmission rate is higher for dual eligible patients, 19.28%, compared with 14.83% for all other patients. The readmission rate for African-American patients was also higher at 19.16% compared with 15.10% for patients of all other races. Similarly, the readmission rate for patients in the lowest SES quartile by AHRQ Index was 16.81% compared with 15.05% for all other patients.  Incremental effect of SES variables and race in a multivariable model  We then examined the strength and significance of the SES variables and race in the context of a multivariable model. When we include any of these variables in a multivariate model that includes all of the claims-based clinical variables, the effect size of each of these variables is small. The c-statistic is essentially unchanged with the addition of any of these variables into the model. Furthermore, 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’ RSRRs with the addition of any of these variables. The median absolute change in hospitals’ RSRRs when adding a dual eligibility indicator is 0.004% (interquartile range [IQR] -0.017% – 0.024%; minimum -0.309% – maximum 0.135%) with a correlation coefficient between RSRRs for each hospital with and without dual eligibility added of 0.99836. The median absolute change in hospitals’ RSRRs when adding a race indicator is 0.011% (IQR -0.010% – 0.033%; minimum -0.671% – maximum 0.130%) with a correlation coefficient between RSRRs for each hospital with and without race added of 0.99814. The median absolute change in hospitals’ RSRRs when adding a low SES AHRQ indicator is 0.007% (IQR -0.033% – 0.036%; minimum -0.322% – maximum 0.135%) with a correlation coefficient between RSRRs for each hospital with and without low SES added of 0.99691.  As an additional step, a decomposition analysis was performed. The results are described in the table below.  The patient-level and hospital-level dual eligible, race, and low AHRQ SES Index effects were significantly associated with each of the hospital wide readmission models (Medicine, Surgery, Cardiorespiratory, Cardiovascular, and Neurology) in the decomposition analysis. If the dual eligible, race, or low AHRQ SES Index variables are used 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.  Given these findings and the complex pathways that could explain any relationship between SES or race with readmission, we did not incorporate SES variables or race into the measure. |

Table 6. HWR Decomposition Analysis

| **Parameter** | **Estimate (Standard Error)** | **P-value** |
| --- | --- | --- |
| **Dual Eligible – Patient-Level – Medicine** | 0.0599 (0.00433) | <.0001 |
| **Dual Eligible – Hospital-Level – Medicine** | 0.3207 (0.0177) | <.0001 |
| **Dual Eligible – Patient-Level – Surgery** | 0.1483 (0.00794) | <.0001 |
| **Dual Eligible – Hospital-Level – Surgery** | 0.4743 (0.0332) | <.0001 |
| **Dual Eligible – Patient-Level – Cardio Respiratory** | 0.1043 (0.00634) | <.0001 |
| **Dual Eligible – Hospital-Level – Cardio Respiratory** | 0.4148 (0.0269) | <.0001 |
| **Dual Eligible – Patient-Level – Cardiovascular** | 0.1607 (0.0101) | <.0001 |
| **Dual Eligible – Hospital-Level – Cardiovascular** | 0.5318 (0.0418) | <.0001 |
| **Dual Eligible – Patient-Level – Neurology** | 0.0874 (0.0129) | <.0001 |
| **Dual Eligible – Hospital-Level – Neurology** | 0.4997 (0.0526) | <.0001 |
| **African American – Patient-Level – Medicine** | 0.0374 (0.00558) | <.0001 |
| **African American – Hospital-Level – Medicine** | 0.3208 (0.0119) | <.0001 |
| **African American – Patient-Level – Surgery** | 0.0959 (0.0103) | <.0001 |
| **African American – Hospital-Level – Surgery** | 0.4423 (0.0214) | <.0001 |
| **African American – Patient-Level – Cardio Respiratory** | 0.0470 (0.00884) | <.0001 |
| **African American – Hospital-Level – Cardio Respiratory** | 0.3386 (0.0186) | <.0001 |
| **African American – Patient-Level – Cardiovascular** | 0.0763 (0.0131) | <.0001 |
| **African American – Hospital-Level – Cardiovascular** | 0.3501 (0.0269) | <.0001 |
| **African American – Patient-Level – Neurology** | 0.1200 (0.0155) | <.0001 |
| **African American – Hospital-Level – Neurology** | 0.5252 (0.0331) | <.0001 |
| **AHRQ SES Index – Patient-Level – Medicine** | 0.0249 (0.00444) | <.0001 |
| **AHRQ SES Index – Hospital-Level – Medicine** | 0.0788 (0.00653) | <.0001 |
| **AHRQ SES Index – Patient-Level – Surgery** | 0.0349 (0.00689) | <.0001 |
| **AHRQ SES Index – Hospital-Level – Surgery** | 0.1254 (0.0120) | <.0001 |
| **AHRQ SES Index – Patient-Level – Cardio Respiratory** | 0.0376 (0.00661) | <.0001 |
| **AHRQ SES Index – Hospital-Level – Cardio Respiratory** | 0.1105 (0.00910) | <.0001 |
| **AHRQ SES Index – Patient-Level – Cardiovascular** | 0.0307 (0.00943) | 0.0011 |
| **AHRQ SES Index – Hospital-Level – Cardiovascular** | 0.1375 (0.0149) | <.0001 |
| **AHRQ SES Index – Patient-Level – Neurology** | 0.0544 (0.0125) | <.0001 |
| **AHRQ SES Index – Hospital-Level – Neurology** | 0.1314 (0.0198) | <.0001 |

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

|  |
| --- |
| We tested the performance of the model across the development, validation, and 2012 Samples (**Dataset 1**). We computed three summary statistics for assessing model performance (Harrell and Shih, 2001):  Discrimination statistics  (1) Area under the receiver operating characteristic (ROC) curve (the c-statistic (also called ROC) 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.)  References:  F.E. Harrell and Y.C.T. Shih, 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*** [***2b4.9***](#question2b49)

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C-statistics and discrimination in the Development, Validation, and 2012 Sample in **Dataset 1** by specialty cohort:  Table 7. C-statistics and discrimination in the Development, Validation, and 2012 Sample (Dataset 1)   |  | Hybrid HWR Measure Development Sample | Hybrid HWR Measure Validation Sample | Hybrid HWR Measure 2012 Sample | | --- | --- | --- | --- | | c-statistics | | | | | Surgery/Gynecology | 0.802 | 0.799 | 0.800 | | Cardiorespiratory | 0.668 | 0.673 | 0.666 | | Cardiovascular | 0.731 | 0.717 | 0.726 | | Neurology | 0.708 | 0.697 | 0.693 | | Medicine | 0.651 | 0.656 | 0.665 | | Discrimination-Predictive Ability (lowest decile %, highest decile %) | | | | | Surgery/Gynecology | 0-35 | 0-36 | 0-31 | | Cardiorespiratory | 9-39 | 7-41 | 6-36 | | Cardiovascular | 2-29 | 2-32 | 2-24 | | Neurology | 4-33 | 5-37 | 5-34 | | Medicine | 8-35 | 7-35 | 6-34 | |

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Calibration in the Development, Validation, and 2012 Sample in **Dataset 1** by specialty cohort:  Table 8. Calibration in the Development, Validation, and 2012 Sample (Dataset 1)   |  | Hybrid HWR Measure Development Sample | Hybrid HWR Measure Validation Sample | Hybrid HWR Measure 2012 Sample | | --- | --- | --- | --- | | Calibration (γ0, γ1) | | | | | Surgery/Gynecology | (0.000, 1.000) | (-0.049,0.948) | (-0.192,0.971) | | Cardiorespiratory | (0.000, 1.000) | (-0.004,0.995) | (-0.111,0.931) | | Cardiovascular | (0.000, 1.000) | (0.067,1.007) | (-0.333,0.854) | | Neurology | (0.000, 1.000) | (-0.129,0.920) | (-0.464,0.781) | | Medicine | (0.000, 1.000) | (-0.047,0.977) | (0.077,1.108) | |

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

|  |
| --- |
| See Calibration Curves in the Excel Data Dictionary attached in data field S.2b. |

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

|  |
| --- |
| N/A This measure is not stratified. |

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

|  |
| --- |
| In **Dataset 1**:  Discrimination Statistics  The range of c-statistics from 0.651 to 0.802 showing good to excellent discrimination across the specialty cohort models.  Calibration Statistics  The calibration value of close to 0 at one end and close to 1 to the other end indicates good calibration of the model. The risk decile plot shows excellent discrimination of the model and good predictive ability.  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. |

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

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

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

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

|  |
| --- |
| The method for discriminating hospital performance has not been determined. For public reporting of measures of hospital outcomes developed with similar methodology, CMS characterizes the uncertainty associated with the RSRR by estimating the 95% interval estimate. This is similar to a 95% confidence interval but is calculated differently. If the RSRR’s interval estimate does not include the national observed mortality rate (is lower or higher than the rate), then CMS is confident that the hospital’s RSRR 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 RSSR 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.  However, the measure is not currently publicly reported, and decisions about the approach to discriminating hospital performance have not been made. |

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

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

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

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

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

**Note***: This criterion is directed 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).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

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

|  |
| --- |
|  |

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

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

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

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

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

|  |
| --- |
| For the EHR data elements used in the measure’s risk models, we anticipate that there will be some missing data. We examined the rates of data capture, and missing data, in the development and testing samples (**Dataset 1**) as well as in the EHR data element feasibility and validity datasets (**Datasets 2** and **3**). We also examined the distribution of the CCDE data values in **Dataset 1** to determine what proportion were out of physiological range and might represent data errors. We found that most values fell within physiological range and that there were few apparent errors in the data entry.  As was shown in 2a.2.3 and 2b.2.3, in all datasets used for testing the rates of capture are above 90% for the data elements included the risk models. Because missing values were rare in the development and testing datasets, it was not necessary to do tests of bias in measure results. However, in order to reduce any small chance of bias due to missing data, we set missing values to the median value in all measure risk models and included a dummy variable whenever a data element was missing in 5% or more of the admissions in each specialty cohort.  To reduce the effect of the spurious outliers, we transformed extreme values by replacing them with a value at the outer limit of a designated range by a process called Winsorization1,2. All continuous variables with values less than 1st percentile or higher than the 99th percentile were Winsorized (i.e., values less than the 1st percentile were assigned to the value of the 1st percentile, and values greater than the 99th percentile were assigned to the value of the 99th percentile). Missing data values were set to the median value for the cohort. In addition, dummy variables for missing data were included in the statistical models.  References:  1. Altenburg HP. Estimation of Radioimmunoassay Data Using Robust Nonlinear Regression Methods. In: Dodge Y, Whittaker J, eds. Computational Statistics: Physica-Verlag HD; 1992:367-372.  2. Dixon WJ, Yuen KK. Trimming and winsorization: A review. Statistische Hefte. 1974/06/01 1974;15(2-3):157-170 |

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

We report the capture rate of all EHR data elements in each dataset in section 2a.2.3, and 2b.2.3.

The range of missing data across hospitals in **Dataset 1** can be found in the Data Dictionary attached in data field S.2b

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

The rate of missing values was low in all of the datasets and for all hospitals used for testing and therefore not likely to introduce bias. However, we did account for potential outlier values as well as missing values in our risk models to reduce any small possibility of bias. However, approaches to handling missing clinical data in measure calculation will be reassessed during implementation.