**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b6)**

**Measure Title**: All-Cause Unplanned Readmission Measure for 30 Days Post Discharge from Inpatient Rehabilitation Facilities

**Date of Submission**: 2/5/2014

**Type of Measure:**

|  |  |
| --- | --- |
| ☐ Composite – ***STOP – use composite testing form*** | X 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.  **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.    **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 decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **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**.  **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 |
| X administrative claims | X administrative claims |
| ☐ clinical database/registry | ☐ clinical database/registry |
| ☐ abstracted from electronic health record | ☐ abstracted from electronic health record |
| ☐ eMeasure (HQMF) implemented in EHRs | ☐ eMeasure (HQMF) implemented in EHRs |
| X other: administrative enrollment/eligibility files | X other: administrative enrollment/eligibility files |

**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*).

* Medicare Inpatient claims - standard analytical files (2007-2012). (Note that the index IRF claims cover the 2009-2011 period; however, the inpatient claims data which identify the prior acute care claim extends back to 2007 in order to include very long lengths of stay.)

Documentation for the Medicare claims data is provided online by the CMS contractor, Research Data Assistance Center (ResDAC) at the University of Minnesota. The following web page includes data dictionaries for these files: Standard analytical files (Inpatient RIF): <http://www.resdac.org/cms-data/files/ip-rif/data-documentation>

* Medicare Enrollment Database

<http://aspe.hhs.gov/datacncl/datadir/cms.htm>

* Medicare Denominator files (2009-2011)

http://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/IdentifiableDataFiles/DenominatorFile.html

* AHRQ CCS groupings of ICD-9 codes

<http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>

* CMS-HCC mappings of ICD-9 codes

Mappings are included in the software at the following website:

<http://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors.html>

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

2009, 2010, and 2011

**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 |
| X hospital/facility/agency | X 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*)

To examine the reliability of this readmission measure, we performed analyses of Medicare claims, Denominator, and EDB files from 2009-2011, identifying IRF admissions preceded by an acute-care hospitalization (IPPS, CAH, and psychiatric hospitals). Because this measure is based on two consecutive years of data, these analyses used data pooled from 2009/2010 and 2010/2011. After applying the exclusion criteria the final analytic files included the following counts of stays and facilities:[[1]](#footnote-1)

2009/2010: 587,810 index IRF stays in 1,185 IRFs

2010/2011: 590,120 index IRF stays in 1,171 IRFs

Next, consistent with the reliability testing done for the Hospital Wide Readmission (HWR) QM (NQF #1789), we conducted an analysis based on a random split of the data to assess reliability and consistency of the measure. We pooled stay-level data from 2009, 2010, and 2011 across all facilities and split the file randomly into two data sets. The two data sets derived from the three years of pooled data were then used for reliability testing. The final analytic files had the following counts of stays and facilities:

Sample 1: 442,555 index IRF stays in 1,206 facilities

Sample 2: 443,493 index IRF stays in 1,206 facilities

Note: The number of facilities included in the split sample files exceeds the number reported above because a small number of facilities are not included in all three years. When the file is randomly split, stays from these facilities are split across both files.

The split samples are each smaller than the samples that would be used in the proposed measure, which have 2 years of data in each measurement period. A fourth recent year was not available to allow split samples each of a size equal to 2 data years.

Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program references: lc22\_irf\_gv150910.xlsx; lc22\_irf\_gv151011.xlsx; lc26\_glim\_irfv150.xlsx; and lc26\_glim\_irfv151.xlsx)

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

This measure is based on the 100% claims data and includes all IRF users during these years at least age 18, who had the Medicare Part A coverage needed and had a prior acute care stay within 30 days. A patient exclusion from the measure is that of cancer patients who were not treated surgically in the prior acute stay. All these and technical exclusions are explained in section S.10 and listed in **Testing Figure 1** below.

**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**.

No differences beyond what is described in 1.5.

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

**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*)  
X **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)   
X **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**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 Item Reliability***

To enhance the reliability of the model, RTI selected data elements considered most robust and reliable from prior research using the source data to build the sample and include in the model. Where possible, we approached variable selection for the development of this measure to be harmonized with the construction of the HWR (NQF #1789). In employing this approach we cite the same justification used for the HWR with regard to reliability of data elements used. Similar to NQF #1789, we selected data elements focusing on variables that are likely to be coded more consistently across hospitals and IRFs because they are used for payment or are audited. For example, consistent with the HWR, we used admission and discharge dates on IRF and hospital claims to identify transfers and readmissions, rather than relying on the claim “discharge status” data elements. (We do use the discharge status value “expired,” however. Death dates in the Medicare eligibility files are not always accurate, especially for beneficiaries eligible through the Railroad Board; these beneficiaries usually have the last day of a month as a recorded death date.) We also note that CMS has an audit process in place for hospitals which includes review of diagnosis and procedure codes.

In addition to our approach to harmonize this measure based on an existing measure and research, we also analyzed the reliability of the data items included in our models. First, we examined the consistency of covariate prevalence across model years (2009/2010 compared to 2010/2011). Next, we assessed the coefficient estimates and precision of these estimates (i.e., confidence intervals) across model years in order to assess model stability. We also assessed the prevalence, estimates, and precision of the covariates estimated from the split sample analysis described below.

***Quality Measure Reliability***

To evaluate the reliability of the quality measure we followed the test-retest approach used in the evaluation of the HWR (NQF #1789), which involved examining the level of agreement between facilities’ scores when calculated based on two mutually exclusive random samples of stays within each facility. We combined the 2009, 2010, and 2011 files and took a random sample of stays, splitting the combined years for each facility into halves with approximately equal numbers from each year. We recalculated the standardized risk ratio (SRR) for each facility based on each data set. The level of agreement between the SRRs calculated on the two different samples gives us a test of the repeatability of the measure. Agreement was evaluated using intraclass correlations (ICC) with the facility as the cluster, calculated assuming a random subset of all possible raters (Shrout and Fleiss 1979). ICCs are reported overall and by quartiles of facility size.

Another test to evaluate model stability is to analyze the correlation among SRRs across the two consecutive measurement periods using a Pearson correlation. We also calculated the correlation coefficient on the original files to assess consistency of facilities’ SRRs across models.

***Reference:***

Shrout PE, and Fleiss JL Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*. 1979, 86, 420-428.

**2a2.3. For each level 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***

We examined the consistency of the prevalence of covariates and odds ratio estimates between the model constructed using 2009/2010 data and the model constructed using 2010/2011 data. **Appendix C** includes the results from both models. The vast majority of covariates had similar prevalence from model to model; for example, 86 percent,176, of the 204 covariates had changes in prevalence of 20 percent or less. Twenty-three covariates (11%) had changes in prevalence between 20-50 percent, 4 (2%) had changes between 50-100 percent, and 1 covariate (the marker for supplemental oxygen) increased by more than 100 percent between 2009/2010 and 2010/2011. Odds ratios for these covariates were very similar between models. We used the 95% confidence intervals for each covariate to assess overlap of the odds ratios and found that there was overlap for all confidence intervals between models.

In summary, the findings from these comparisons suggest that there were some changes in the prevalence rates over time, but there were no significant differences in the estimated coefficients for the covariates between the models.

We also compared the consistency of covariate prevalence and estimates for the two split sample files. The two models based on the split sample were internally similar. Specifically, the prevalence of all 204 covariates included in the first split sample were within 7.5 percent of the covariates in the second split sample. Odds ratios estimated from each split sample model showed similar consistency: 95% confidence intervals overlapped for all but two of the covariates. The HCC groups for diabetes and neuropathy were the only covariates to show a significant difference by comparison of the 95% confidence intervals. These differences are likely due to random variation.

***Quality Measure Reliability***

Examining the level of agreement between SRR scores calculated on each of the split files, we found an ICC of 0.39, indicating a modest level of agreement between facilities’ SRRs. When stratified by quartile of IRF count of stays across the three years, the observed ICCs on the split sample comparison were as reported in the **Testing Table 1** below:

**Testing Table 1: Reliability Analysis: Intra-Class Correlations (ICC) for IRFs based on Facility Size Quartiles, 2009-2011**

|  |  |  |
| --- | --- | --- |
| **Quartiles: Facility sample size in each subsample – approximate** | **Number of IRFs** | **ICC** |
| Quartile 1: 1-141 | 301 | 0.24 |
| Quartile 2: 142-248 | 300 | 0.18 |
| Quartile 3: 249-460 | 304 | 0.36 |
| Quartile 4: 462-5,066 | 301 | 0.54 |

Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program references: lc26\_corrirfv15\_IRF\_SRR.xls; lc26\_corrirfq15\_.lst; lc26\_corrirfq15\_IRF\_SRR\_quart1.xlsx; lc26\_corrirfq15\_IRF\_SRR\_quart2.xlsx; lc26\_corrirfq15\_IRF\_SRR\_quart3.xlsx; lc26\_corrirfq15\_IRF\_SRR\_quart4.xlsx)

These results indicate that as facility size increases, the level of agreement among SRRs is more stable. We investigated the degree of similarity of patients in the two samples at the facility level. With small samples there is a strong likelihood that the two groups could look quite different. We computed the difference of the mean expected probabilities of readmission for each facility’s split samples. The mean of absolute values of the differences was .0025 for facilities with more than 1000 stays, 0.0053 for facilities with 200 – 400 stays and 0.0208 for facilities with less than 55 stays. This measure of the patient profile shows large differences in patient characteristics when the test is made with small samples. The 2 percentage point difference in risk adjusted probabilities for the small facilities is quite large, 10 times that found in the large facilities.

Our next reliability test to evaluate model stability calculates the Pearson correlation between SRRs across the two consecutive measurement periods. We found a correlation coefficient of 0.70 (p<0.001) which suggests very high correlation between facilities’ SRR scores across measurement periods. After stratifying the sample based on quartiles of facility size, similar to the approach above, we found Pearson correlation coefficients of 0.63, 0.65, 0.68, and 0.77 (all p<0.001) for quartiles 1, 2, 3, and 4, respectively. This suggests high correlations across measurement periods, and greater consistency in the measure as facility size increases.

Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program reference: lc37)

**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?*)  
In evaluating the results of the split samples we have to consider that a smaller than the 2-year required sample size was used. This affected results for all facility sizes. Also the smallest facilities had very different population profiles in the two splits and that reduced the correlations. The larger facilities did have more reasonable correlations. When looking at the measure as it is to be used, over consecutive years with a full 2-year sample, the correlation is much better.

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

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

**2b2.2. For each level 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)*

The focus of the measure is readmission rates for fee-for-service Medicare patients after discharge from an IRF within a window of 30 days. As discussed earlier, there is evidence from short-term hospital studies that readmission rates can be reduced by improved discharge planning and transitions. A risk adjusted measure of the rate will provide information not available to providers currently. Medicare administrative data and eligibility data are used to define the denominator, the readmission and the risk adjusters.

There are patients not included in the measure because they are not FFS Medicare. This group includes patients covered by other payers and those in Medicare MA plans. There is insufficient data on these patients to include them in this measure. Neither their IRF claims nor acute hospital claims are currently available. Whether these patients differ in care and discharge patterns is not known. Payment methods will differ from FFS Medicare and from one another and payment can influence patient care patterns.

Starting with the FFS population, many exclusions (see above) cover stays with anomalous data, patients without sufficient time in Medicare to have a full look-back period, pediatric patients etc. Another group of exclusions is the set of patients who had not been admitted to a short-term acute care or psychiatric hospital within 30 days prior to the IRF admission. This omission, applied to all facilities, might change the absolute value of the readmission rates, but the readmission rates do not have an absolute target. The measure is designed to look at variation around the mean for the population included. About 90% of FFS Medicare patients are included in the measure. The FFS Medicare patient population is both a large proportion of IRF patients and a group of interest in their own right. We expect that Medicare beneficiaries in Medicare Advantage private plans will be includable in the measure when the encounter data submitted by plans are reliably collected.

As an outcome measure, readmissions have been considered to have value applied to other settings and patient groups. Our technical expert panels, including industry representatives and researchers are in agreement with the approach. Validity is partially tested by statistical tests of the model on multiple years of data to predict readmissions and through the assistance of a Technical Expert Panel. The risk adjusters are of the type used to predict other measures of utilization (e.g. hospitalizations), Medicare, Medicaid and private payer spending for medical services, and mortality. The spending models are used by the Federal and State governments to determine payments. The model structure and many of the variables are similar to those in the Hospital Wide All-Cause Readmission measure approved by NQF (#1789).

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

With regard to the ICD-10 transition for the data elements currently specified using ICD-9 codes, our goal was to convert this measure to a new code set, fully consistent with the intent of the original measure. AHRQ has posted a beta version of the mapping between ICD-10 procedure codes and the CCS codes on their website (<http://www.hcup-us.ahrq.gov/toolssoftware/beta/icd_10_beta.jsp>). We plan to use the same CCS groupings in our models after the transition to ICD-10. The final grouper is expected in October 2014. We will continue to monitor and review these mappings of CCS codes to ICD-10 in order to identify any potential changes that may impact this measure. The CMS contractor for the HCCs is currently finalizing the ICD-10 mapping into the HCCs. We plan to use the same set of HCCs, and will review the mapping to ensure that there are no changes that impact this measure. The definition of planned readmissions is currently defined by ICD-9 procedure and diagnosis codes grouped by the Clinical Classification Software (CCS), developed by the AHRQ, where large clusters were appropriate and by individual codes, if necessary. We are awaiting the ICD-10 versions of the HWR planned readmissions codes and we have provided a provisional mapping of these ICD-9s to ICD-10s in **Appendix Table A6**. This mapping was based on the General Equivalence Mappings made available by CMS at <http://www.cms.gov/Medicare/Coding/ICD10/2014-ICD-10-PCS.html>.

**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?*)  
Not applicable.

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

**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*)  
Because the measure exclusions on the FFS sample are largely based on data limitations that would prevent proper analysis, no further analyses were conducted to assess the impacts of these exclusions. In many cases, the lack of available claims data means that further analysis is not feasible. The exclusion of the group of patients who had non-surgical cancer treatment in the prior acute stay was based on the work done in the developing the Hospital-Wide Readmission measure (NQF #1789). Their post-discharge trajectory of readmissions was not consistent with other patient groups.

**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*)

**Testing Figure 1** below summarizes the frequency of exclusions from the denominator of this measure. This flowchart illustrates the various steps in which observations were excluded. As noted, the majority of the exclusions are due to data limitations; for example, nearly 33,000 IRF stays were excluded due to no prior acute claim. Note that the following exclusions were made on all inpatient records, and prior to identifying the IRF claims: excluded claims with no payment information; excluded claims from HMO stays; excluded claims where the patient left against medical advice (AMA); and excluded federal hospital transfers.

Testing Figure 1:    Flowchart of measure exclusions, 2011



Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program references: lc30irf2011v15.doc)

**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*)

As noted, the majority of the exclusions are due to data limitations. Given that this measure utilizes administrative claims data, we have no concerns about missing data distorting provider performance.

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

**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**

X **Statistical risk model with** 204 **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*)

To develop the risk adjustment model for this measure, we analyzed Medicare claims, Denominator, and EDB files for 2009, 2010, and 2011, identifying IRF admissions preceded by an acute-care hospitalization (IPPS, CAH, and psychiatric hospitals). Because this measure is based on 2 consecutive years of data, these analyses used data pooled from 2009/2010 and 2010/2011. After applying the exclusion criteria, as described in Section 2b3, the final analytic files included the following counts of stays and facilities:

2009/2010: 587,810 index IRF stays in 1,185 IRFs

2010/2011: 590,120 index IRF stays in 1,171 IRFs

##### Analytic method

This measure estimates the risk-standardized rate of unplanned, all-cause readmissions for patients (Medicare FFS beneficiaries) discharged from an Inpatient Rehabilitation Facilities (IRF) who were readmitted to a short-stay acute-care hospital or a Long-Term Care Hospital (LTCH), within 30 days of an IRF discharge. (Because an acute readmission on the day after discharge is treated as a transfer, the 30 day count starts on the second day after discharge.) Excluded from the measure (numerator and denominator) are patients who died during the IRF stay; patients less than 18 years old; patients who were transferred at the end of a stay to another IRF or short-term acute hospital; patients who were not continuously enrolled in Part A fee-for-service (FFS) Medicare for the 12 months prior to the IRF admission date, and at least 30 days after IRF stay discharge date; patients who did not have a short-term acute-care stay within 30 days prior to an IRF stay admission date; patients discharged against medical advice (AMA); patients for whom the prior short-term acute-care stay was for nonsurgical treatment of cancer; and IRF stays with data needed for modeling that are problematic.

In this section we detail the analytic methods used including the steps completed to develop our final risk adjustment model.

***Covariate Selection – Conceptual Rationale***

The risk adjustment model for this measure accounts for variation across facilities in patient characteristics predictive of readmission using hierarchical logistic regression. The goal of risk adjustment is to account for differences across facilities in patient demographic and clinical characteristics that might be related to the outcome but pre-exist the admission to the IRF.[[2]](#footnote-2) For this reason, patient acuity (case mix) is taken into account by including patients’ clinical status including hospital principal diagnosis and comorbidities in the predictive models. In addition, we included demographic variables (age and sex), and other health service factors such as length of stay during the patient’s prior proximal hospitalization and number of prior hospitalizations in the previous 365 days.

The specific types of risk adjustment variables include the following:

* Age/sex categories.
* Original reason for entitlement being disability.
* Surgery category if present (e.g., cardiothoracic, orthopedic), defined as in the HWR model; the procedures are grouped using the CCS for ICD-9 procedures developed by AHRQ.
* Receiving dialysis in prior acute stay, defined by presence of revenue code.
* Principal diagnosis on prior acute bill (as in the HWR measure, they are grouped clinically using the CCS for ICD-9 diagnoses developed by AHRQ).
* IRF Case-mix groups on the IRF bill.
* Comorbidities from secondary diagnoses on the prior acute bill and diagnoses from earlier acute stays up to 1 year before IRF admission (these are clustered using the Hierarchical Condition Categories [HCC] groups used by CMS). Some of the HCCs are taken only from the prior acute claim and some are taken from all of the prior acute stays in the 1 year lookback.
* Length of stay and length of stay in intensive care in the prior acute hospital stay (categorical to allow the effect of an additional day to differ for shorter and longer stays).
* Counts of prior acute discharges in the 365 days before the IRF admission (categorical).

Variables such as sex, race, or ethnicity are often not included as adjustment variables in models because the standards of care should not vary across demographic markers for vulnerability to disparities in health outcomes and receipt of quality care. However, for some outcomes, an argument can be made that some of these markers (sex and age) are also associated with demonstrated clinical/physiologic differences that can determine risk at the time the patient enters the IRF. However, our analyses did not indicate any systematic patterns in the readmission risk by sex.

To capture comorbidities, we used the secondary medical diagnoses listed on the patient’s prior proximal hospital claim and all diagnoses listed on acute care hospitalizations that occurred in the prior 12 months. We classified these comorbidities using the hierarchical condition categories (HCCs) that RTI developed for CMS (Pope et al., 2000). The HCCs were developed by grouping the 15,000+ ICD-9-CM codes into approximately 800 diagnosis groupings that were then grouped into roughly 200 hierarchical condition categories. The categories are based on clinical and Medicare cost criteria.

Though we find some small differences by whether an IRF was freestanding or a hospital-based unit, we did not adjust for these facility characteristics in the model. The standard of care should not depend on variables such as facility type or size.

***Specific approach to case mix adjustment using the comorbid risk variables***

Our initial risk adjustment models were developed using logistic regression, reducing the need for greater computational intensity of hierarchical modeling while model building. We developed the initial model using 2009/2010 claims data and then refined this model using 2010/ 2011 data.

Our selection of comorbid risk variables differs from the process used for the HWR, which built on previous work done for the development of condition-specific hospital readmission measures. As the HWR population is different from the IRF population, this necessitates different approaches to stratification, risk adjustment, and the exclusion of planned readmissions; however, the overall analytic approach was harmonized to a great degree. The HWR measure created cohorts based on the principal diagnosis, which correspond to hospital care teams. This allowed the common set of risk adjusters to vary in coefficient magnitudes among the cohorts. However, the cohort-based approach is not feasible for this setting due to substantially smaller samples sizes. In addition the TEP members generally agreed that the stratification was not necessary. The comorbidities for this measure are either secondary diagnoses coded for the prior proximal hospitalization or diagnoses from hospitalizations that occurred in the 12 months prior to the index IRF stay, to adjust for patient morbidity beyond the primary condition. The decision to use the comorbidities based on the prior proximal hospitalization versus the 365-day lookback window was based on statistical coefficient results and clinical input in terms of which types of comorbidities are acute exacerbations and more temporary versus conditions that are more chronic and long-term.

Below, we describe the steps we employed for variable selection and development with regard to principal diagnoses (measured using AHRQ CCS) and comorbidities (measures using HCCs), which were included in our final risk adjustment model.

***Principal diagnosis***

1. Initially we ran a logistic regression model that included all of the AHRQ CCS categories along with the demographic and other clinical covariates listed above, and all individual HCCs.
2. This initial model kept all CCS categories ungrouped in the model, but we noted that many CCS categories had very low prevalence in our patient population, and individually these diagnoses were not adding to model prediction. To address these issues, we aggregated these codes into 61 larger groupings of CCS codes. These groupings were based on an extensive expert clinical review process of identifying conditions that were similar; for example, several of the circulatory system-related CCS codes were combined given their clinical similarities. The reference group chosen for these principal diagnostic categories was a broad grouping of several CCS categories that were all associated with low odds of readmission.

***Comorbidities***

1. We ran the model controlling for the demographic, clinical factors, including the groupings of CCS codes, and evaluated the HCCs individually. We reviewed the coefficients and p-values across the model years (2009/2010 and 2010/2011) for each of the HCCs to determine whether to include the individual HCC, a larger cluster of HCCs, or to exclude the HCC from the final model giving consideration to the consistency of effects across the models and the number of patients with the comorbidity. With some exceptions, we selected the final set of HCC variables based on the following general principles:

i. We excluded HCCs that were not consistently significant or approaching significance in at least one or more of the three years (p<0.10).

ii. We excluded HCC groupings that were predominantly protective or that may have reflected coding practices rather than a patient’s clinical status. We noted that certain comorbidities appeared to be protective because they may have been coded more often for healthier patients with fewer of the more severe comorbidities than for sicker patients who had more competing comorbidities to include in the billing form.

Our final review model includes 59 individual HCCs and 10 groupings of HCCs. Please refer to **Appendix D** for documentation on the final HCC groupings. The majority, 40 of the HCCs, were based on the prior acute diagnoses and the remaining HCCs, 29, were based on diagnoses obtained from the 365-day lookback window.

***Case Mix Groups (CMGs)***

Unique to this readmission measure is the inclusion of the IRF case-mix groups (CMGs), which are assigned to all patients in connection with the IRF payment system. CMGs are mutually exclusive groupings built on the patient’s underlying medical problem related to IRF care, functional motor score, and for some CMGs, cognitive score and age. This variable adds information beyond the acute-care diagnosis. We initially included all CMGs in the models, but ultimately combined CMGs based on sample size, clinical similarity or similar coefficient estimates in the model. Our final model includes 29 groupings of the CMGs. The reference group for the CMGs includes two larger groupings with low marginal probability of readmission, including fracture of lower extremity and replacement of lower extremity joint. The CMGs define the type of IRF patient and capture information on patients for whom the need for rehabilitation is not related to the principal diagnosis in the prior stay.

***Other variables included***

Markers for having had surgery in the proximal acute stay were tested and found to indicate lower probabilities of readmission than the reference of nonsurgical care. An indicator for having disability as the original reason for Medicare eligibility increased the readmission probability. We used hospital claims to detect patients who had dialysis in the facility but who were not in ESRD and also captured ESRD in the comorbidity variables. As expected, both dialysis and ESRD increased readmission probability.

We tested and included other variables that were potentially associated with probability of readmission and might capture information that the demographics, diagnoses, procedures and CMGs do not. Information about severity of a condition, for example, is potentially enhanced by other inpatient claims information.

Other variables we tested included length of stay (LOS) in the prior proximal short stay, LOS in intensive/critical care, and frequency of admissions in the prior year. All of these are positive predictors of readmissions and were included as categorical variables, each covering a range of values to capture nonlinearities and protect against distortions related to outlier values. The prior admission counts capture patterns of hospital use that pre-date the current IRF stay. As these may indicate characteristics such as poor patient compliance, poor living conditions, or lack of access, a facility can be credited if its discharge and transition practices reduce the probabilities of repeat stays.

***Statistical Approach to Measure Development***

Our approach aligns with the CMS HWR measure design. As such, we use some of the same text to explain our approach (Horwitz et al. 2011).

***Models for each patient’s risk of readmission***

For model development, we used logistic regression models with a logistic link function, with outcome Yi for the ith patient equal to 1 if the patient was readmitted within 30 days of discharge and 0 otherwise.

For our final models we extended the logistic regression models to include an additional error term. That is, due to the natural clustering of observations within facilities, we used hierarchical logistic regression to model the log-odds of readmission for each index IRF stay. Readmission within 30 days was modeled as a function of patient-level demographic and clinical characteristics and as a random IRF-level intercept. This model specification accounts for within-IRF correlation of the observed outcomes and the estimates of the IRF effects capture the facility-specific contribution to the readmission rates, after risk adjustment for the patient characteristics.

Specifically, we estimated a hierarchical logistic regression model as follows. In the following, the subscript i indicates a particular patient and subscript j indicates a particular IRF. Let Yij denote the outcome (equal to 1 if patienti is readmitted within 30 days, 0 otherwise) for a patienti at IRFj; *Z*ijdenotes a set of risk factors. We assume the outcome is related linearly to the covariates via a logit function with dispersion:

logit(Prob(Yij =1)) = αj + β\*Zij + εij (1)

αj = μ + ωj ;  ωj ~ N(0, τ2)

where Z ij *=* (Z1, Z2, ... Zk) is a set of *k* patient-level covariates. Each αj represents an IRF-specific intercept; μis the adjusted average outcome over all IRFs; and τ2 is the between IRF variance component and ε ~N(0,σ2) captures any over- or under-dispersion. The hierarchical logistic regression model is estimated using the SAS software (SAS GLIMMIX: SAS/STAT User’s Guide, SAS Institute Inc.)

***IRF performance reporting***

The previous section described how the risk adjustment model was specified and estimated, using a hierarchical logistic regression model. The model was used to calculate a standardized risk ratio (SRR) for each IRF.

We used the results from the hierarchical logistic regression model to calculate the predicted and the expected number of readmissions for each IRF. The predicted number of readmissions for each IRF was calculated as the sum of the predicted probability of readmission for each patient in the facility, including the IRF-specific (random) effect. Using the notation of the previous section, the model specific risk standardized readmission ratio for each IRF is calculated as follows. To calculate the predicted number of readmissions predj for index IRF stays at IRFj, we used

predj = Σlogit-1(μ + ωj + β\*Zij) (2)

where the sum is over all stays in IRFj, and ωi is the IRF intercept. To calculate the expected number expj we used

expj = Σlogit-1 (μ + β\*Zij) (3)

in which the IRF-specific effect is removed. Then, as a measure of excess or reduced readmissions among index stays at IRFj, we calculated the standardized risk ratio SRRj as

SRRj = predj/expj (4)

***Risk-standardized IRF 30-day readmission rate***

This value, SRRj, is the standardized risk ratio for IRFj. To aid interpretation, the standardized risk ratio, SRRj, is then multiplied by the overall national raw readmission rate for all IRF stays, *Ῡ*, to produce the risk-standardized readmission rate (RSRRj).

RSRRj = SRRj\*Ῡ (5)

NOTE: Because the statistic described in Equation (5) is a complex function of parameter estimates, re-sampling and simulation techniques (e.g. bootstrapping) are necessary to derive an interval estimate for the final risk-standardized rate to characterize the uncertainty of the estimate.

***References:***

Horwitz L, Partovian C, Lin Z, Herrin Jeph, Grady J, Conover M, Montague J, Dillaway C, Bartczak K, Suter L, Ross J, Bernheim S, Krumholz H, Drye E. Hospital-Wide All-Cause Unplanned Readmission Measure Final Technical Report. July 2012.

Pope GC, Ellis RP, Ash AS, et al. Principal inpatient diagnostic cost group model for Medicare risk adjustment. *Health Care Financing Review*. 2000; 21(3):93-118.

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

We fit a logistic regression model to estimate readmission rates. The final set of risk variables included in each model, along with the percentage of patients with the characteristic of each risk variable and the adjusted odds ratio for each risk variable, are detailed in **Appendix C**.

The model yielded an overall C-statistic of 0.69, which is equal to or greater than that from other comparable readmission measures.

We estimated a hierarchical logistic regression model and calculated the respective RSRR for each facility. **Testing Table 2** below shows the distribution across facilities of the unadjusted (observed) and risk standardized readmission rates from the model.

As shown in **Testing Table 2**, the unadjusted unplanned readmission rates range from 0.0% to 100.0%, with a median of 12.9% and an interquartile range of 10.7%-15.1%. In contrast, the RSRR has a much narrower range, from 11.1% to 16.1%, with a slightly higher median of 13.5% and a tighter interquartile range of 13.0%-13.9%. The mean RSRR (13.5%) is also slightly higher than the unadjusted rate (13.1%) and the RSRR scores have a much smaller standard deviation (0.7% vs. 4.6% for the unadjusted unplanned readmission rate).

Testing Table 2: Distribution of unadjusted and risk-standardized readmission rates among IRFs, 2010/2011

|  | **N** | | **Mean** | | **SD** | **Min** | **10th Pctl** | **25th Pctl** | **Median** | **75th Pctl** | **90th Pctl** | **Max** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Unadjusted readmission rate** | | 1,171 | 13.1 | 4.6 | | 0.0 | 8.7 | 10.7 | 12.9 | 15.1 | 17.2 | 100.0 |
| **Risk-standardized readmission rate** | | 1,171 | 13.5 | 0.7 | | 11.1 | 12.6 | 13.0 | 13.5 | 13.9 | 14.4 | 16.1 |

Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program references: lc32\_RSRR\_IRF\_1011.xlsx)

The distributions of the unadjusted and IRF-level risk-standardized readmission rates (RSRR) are also illustrated in **Testing Figures 2 and 3**, respectively, where the vertical axis indicates the percentage of IRFs and the horizontal axis the readmission rate. Note that the range on the x axis differs between the two figures and the distribution of the RSRRs is much narrower than the raw rates.

Testing Figure 2: Distribution of unadjusted readmission rates among IRFs

[N=1,171; Mean (StD) 13.1 (4.6)]

Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program reference: lc22\_irf\_gv15\_ALL\_rsrr\_1011.xlsx)

Testing Figure 3: Distribution of risk standardized readmission rates (RSRR) among IRFs

[N=1,171; Mean (StD) 13.5 (0.1)]

Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program reference: lc22\_irf\_gv15\_ALL\_rsrr\_1011.xlsx)

**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*)

*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)

In summary, these results demonstrate good model fit and reasonable predictive ability of our risk adjustment models in profiling IRFs by the measure of risk standardized 30-day readmission rate. Further justification for our risk adjustment model can be seen from **Testing Table 3** which provides calibration results for the two model years of data we analyzed: 2009/2010 and 2010/2011. Using logistic regression results, we computed five summary statistics to assess model performance: calibration, discrimination in terms of predictive ability, discrimination in terms of the C-statistic (equivalent to area under the receiver operating characteristic curve [ROC]), distribution of residuals, and model chi‐square.

The calibration test regresses, using a logit equation, the actual dependent variable on the predicted value. The parameters γ0 and γ1 are the intercept and slope coefficients. The former should be close to zero and the latter close to one. This is a test of the average calibration of the model and, when applied to the development sample, should be passed. Our statistics match these expectations. Because the nature of this measure is to re-estimate the parameters on each sample to which it is applied, average calibration should be good.

Discrimination in predictive ability assesses the ability to distinguish high-risk from low-risk subjects. Each year’s model demonstrates good discrimination, as **Testing Table 3** shows in each case there is a wide range between the mean predictive probability in the lowest decile versus the highest decile.

The C-statistic is a measure of how accurately a statistical model is able to distinguish between a patient with and without an outcome. For binary outcomes the C-statistic is identical to the area under an ROC curve for the model. A C-statistic of 0.50 indicates random prediction, implying the model predicts no better than random chance. A C-statistic of 1.0 indicates perfect prediction, implying the model is perfectly predictive of the outcome. In these models, the range of C-statistic results is consistently around 0.69, which is in line with observed results for other 30-day readmission measures. **Appendix B** includes the ROC curves for these models.

The distribution of the Pearson residuals falling in range shows results very similar to the HWR models that Yale developed. Finally, the Likelihood Ratio model chi-squares show the overall model fit in the two measurement periods. These summary statistics provide further justification for the fit and predictive ability of our risk adjustment model in profiling IRFs by the measure of risk standardized 30-day readmission rate.

Testing Table 3: Model calibration results for 2009/2010 and 2010/2011 analytic files created from the Medicare claims data

| **Indices** | **2009/2010** | **2010/2011** |
| --- | --- | --- |
| Calibration (γ0, γ1) from regression readmission = γ0 + γ1 \* predicted | 0,1 | 0,1 |
| Discrimination - Predictive Ability (lowest decile mean %, highest decile mean %) | 4.4%, 32.2% | 4.3%, 33.1% |
| Discrimination - C-statistic | 0.689 | 0.694 |
| Distribution of residuals (% stays with Pearson Residual Falling in range) | | |
| <-2 | 0.0 | 0.0 |
| -2 to <0 | 86.5 | 86.5 |
| 0 to<2 | 4.7 | 4.9 |
| >2 | 8.8 | 8.6 |
| Model x2 (DF) from Likelihood Ratio | 30974.187 (204) | 32949.772 (204) |

Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program references: lc18\_irfv15\_0910logistic.xlsx; lc18\_irfv15\_1011logistic.xlsx; lc17\_irfv15\_0910.xlsx; lc17\_irfv15\_1011.xlsx; lc19\_irf\_irf0910logistic.xlsx; lc19\_irf\_irf1011logistic.xlsx).

A test that explores calibration over ranges of predicted probabilities is a comparison of the observed and predicted readmissions by decile. Results from this test are reported in **Testing Table 4**. The results of this test are similar across years, and only results from the 2010/2011 model are reported. These results indicate that the difference between the predicted number of readmissions and the observed number of readmissions in percentage points is very minimal, 1 percentage point or less for 9 deciles and less than 2 points for the tenth.

**Testing Table 4: IRF Readmission Model Diagnostics: Comparison of Observed and Predicted Readmissions by Expected Readmission Deciles – 2010/2011**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Decile based on Expected**  **(Low to high)** | **Number of IRF Discharges** | **Number of Observed**  **Readmissions** | **Number of Predicted**  **Readmissions** | **Difference: Predicted – Observed (% points)** |
| **1** | 59,012 | 2,088 | 2,546.22 | 0.78 |
| **2** | 59,012 | 3,116 | 3,592.36 | 0.81 |
| **3** | 59,012 | 4,175 | 4,402.25 | 0.39 |
| **4** | 59,012 | 5,135 | 5,198.58 | 0.11 |
| **5** | 59,012 | 6,375 | 6,082.38 | -0.50 |
| **6** | 59,012 | 7,510 | 7,107.43 | -0.68 |
| **7** | 59,012 | 8,954 | 8,363.64 | -1.00 |
| **8** | 59,012 | 10,470 | 10,045.73 | -0.72 |
| **9** | 59,012 | 13,170 | 12,676.37 | -0.84 |
| **10** | 59,012 | 18,560 | 19,538.05 | 1.66 |

Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program reference: lc17\_irfv15\_1011.xlsx).

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**See Testing Table 3 above.

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
See Testing Table 3 above.

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
See 2b4.5 and Appendix B for ROC Curves.

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

N/A

**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 summary, these results demonstrate good model fit and reasonable predictive ability of our risk adjustment models in profiling IRFs by the measure of risk standardized 30-day readmission rate.

\***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*)

None

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

**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)*   
We used hierarchical, multivariate risk-adjustment models to derive the facility-level 30-day readmission rate. The measure is not an estimate based on samples; rather it includes all IRF patients nationwide who meet the inclusion criteria.

However, the method for determining meaningful differences in facility performance has not yet been determined. For several publicly reported readmission measures of hospital outcomes developed with similar methodology, CMS currently estimates an interval estimate for each risk-standardized rate to characterize the amount of uncertainty associated with the rate (compares the interval estimate to the national crude rate for the outcome), and categorizes hospitals as “better than,” “worse than,” or “no different than” the US national rate. However, the decision to publicly report this measure and the approach to discriminating performance has not been determined.

For our measure development, we conducted bootstrapping in order to estimate multiple estimates of IRFs’ RSRRs. This approach allows for the estimation of confidence intervals around the RSRRs, and for the identification of readmission rates relative to the average facility’s performance. We ran 1,000 bootstrap samples of facilities for this analysis.

**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*)

The distribution of the Risk Standardized Readmission Rate (RSRR) is shown above in **Testing Table 2**. The RSRR has a mean of 13.5% (SD: 0.7%) and a range from 11.1% to 16.1%, with a median of 13.5% and an interquartile range of 13.0-13.9%. The distribution of RSRRs is roughly bell-shaped (**Testing Figure 3**), as would be expected with a random effects model that assumes a normal distribution for the facility effects. (The distribution of unadjusted rates was similar with some cases at the extremes of 0 and 100 percent readmissions.)

In order to identify meaningful differences in performance between providers, we conducted bootstrapping to estimate confidence intervals around the RSRRs allowing for comparison between providers and the national average. These results are summarized in **Testing Table 5** below. We found that 74% of IRFs overall were significantly different than the national average readmission rate. The percent significantly different increases as facility size increases; for example, 41% of IRFs in the smallest decile based on volume were significantly different compared to 82 percent significantly different in decile 10, the largest IRFs.

The last two columns present the percent that are significantly higher (worse) and significantly lower (better) than average. Across all deciles, the proportion significantly higher and the proportion significantly lower are roughly the same. However, within some deciles, the proportion higher and lower is less balanced. For example, in decile 6, 44% of the providers are significantly higher or worse than average compared to less than 35% that are lower or better than average.

**Testing Table 5: Summary Results of Bootstrapping, IRF 2010/2011**

**Percent of IRFs Statistically Significantly Different from National Mean, Overall and by Deciles**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Volume Deciles | Number of facilities | Number significantly different | Percent significantly different | Percent significantly higher (worse) | Percent significantly lower (better) |
| Decile 1 | 116 | 47 | 40.5% | 16.4% | 24.1% |
| Decile 2 | 116 | 87 | 75.0% | 39.7% | 35.3% |
| Decile 3 | 120 | 85 | 70.8% | 34.2% | 36.7% |
| Decile 4 | 117 | 82 | 70.1% | 30.8% | 39.3% |
| Decile 5 | 118 | 92 | 78.0% | 38.1% | 39.8% |
| Decile 6 | 116 | 91 | 78.4% | 44.0% | 34.5% |
| Decile 7 | 117 | 92 | 78.6% | 35.9% | 42.7% |
| Decile 8 | 117 | 101 | 86.3% | 46.2% | 40.2% |
| Decile 9 | 117 | 96 | 82.1% | 42.7% | 39.3% |
| Decile 10 | 117 | 96 | 82.1% | 37.6% | 44.4% |
| *Overall* | 1,171 | 869 | 74.2% | 36.5% | 37.7% |

Source: RTI analysis of Medicare claims data, 2007-2012. (RTI program reference: mi16\_irf\_tab.xlsx and mi16\_irf\_tab2.xlsx).

Note: Deciles based on provider stay count for 2010/2011.

**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?*)

Though the policy decision has not yet been determined by CMS in terms of how IRFs readmission rates may be reported with respect to IRFs nationally, results of the bootstrapping analyses suggest the ability to discriminate between providers’ performance for this readmission measure. The fact that the percent significantly different from the national average is lower among IRFs with fewer stays suggests that public reporting might incorporate a minimum reporting threshold; however, this policy decision has not been determined.

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

**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 datasources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)  
N/A

**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*)  
N/A

**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*)  
N/A

1. The facility counts and analyses presented include both freestanding IRFs and hospital-based IRF units. Also, note that facilities could be in any or all of the years. [↑](#footnote-ref-1)
2. The CMG variables are an exception, see below. [↑](#footnote-ref-2)