NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Standardized Readmission Ratio for dialysis facilities

Date of Submission: 2/5/2014

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 <u>all</u> measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For <u>outcome and resource use</u> measures**, section **2b4** also must be completed.
- If specified for **<u>multiple data sources/sets of specificaitons</u>** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to <u>all</u> 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.

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:	☐ other:

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

These data are part of an extensive and comprehensive national ESRD patient database, derived from Program Medical Management and Information System (PMMIS/REMIS), Medicare claims, the Standard Information Management System (SIMS) database maintained by the 18 ESRD Networks, the CMS Annual Facility Survey (CMS Form 2744), the CMS Medical Evidence Form (CMS Form 2728), the Death Notification Form (CMS Form 2746), and the Social Security Death Master File.

1.3. What are the dates of the data used in testing? 1/1/2009-12/31/2009 for index discharges and 1/1/2009-1/30/2010 for readmissions.

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:	☐ other:

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*)

We included all Medicare-certified facilities treating Medicare dialysis patients (*n* = 6,127) in 2009.

Median facility size was 83 patients. Most facilities were free-standing (80.4%) and located in non-rural areas (79.4%).

1.6. How many and which <u>patients</u> 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*)
After applying the exclusion criteria, these data represent all Medicare-paid hospital discharges for Medicare dialysis patients (*n* = 234,717) during 2009. Patients were mostly white (59.4%) and male (52.9%); the most common types of diagnoses were Complications of Device, Implant or Graft (CCS 237); Congestive Heart Failure, Nonhypertensive (CCS 108); and Hypertension with Complications and Secondary Hypertension (CCS 99).

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.
Not applicable.

2a2. RELIABILITY TESTING
***<u>Note</u>**: 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*)
If the measure were a simple average across individuals in the facility, the NQF-recommended approach for determining measure reliability would be a one-way analysis of variance (ANOVA), in which the between and within facility variation in the measure is determined.1 The inter-unit reliability (IUR) measures the proportion of the measure variability that is attributable to the between-facility variance. The SRR, however, is not a simple average and we instead estimate the IUR using a bootstrap approach, which uses a resampling scheme to estimate the within facility variation that cannot be directly estimated by ANOVA. Refer to the Appendix for a detailed description of this methodology.

See also 2b2 for validity testing of data elements.

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*)

Overall, we found that IUR = .55 (F statistic = 2.24), which indicates that about one half of the variation in the SRR can be attributed to the between-facility differences and about half to within-facility variation.

Table 1. Inter-unit Reliability Measure of SRR, by Facility Size (2009)

Facility Size (No. of Patients)	No. of Facilities	IUR	F-statistic
Small (<=70)	1732	.46	1.85
Medium (71–121)	1784	.54	2.18
Large (>121)	1757	.61	2.53

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?*)
This value of IUR indicates a moderate degree of reliability. When stratified by facility size, we find that, as expected, larger facilities have greater IUR.

____
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 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)*

Validation of critical data elements: The critical data elements for this measure (hospital admission and discharge dates for Medicare dialysis patients) come from Medicare claims data. The validity of these data is ensured by the oversight of the Medicare program in the payment process.

Validation of performance measure score: We assessed the validity of the measure through various comparisons of this measure with other quality measures in use, and in May 2012, presented a preliminary version of the SRR to a CMS Technical Expert Panel (TEP) for clinical validity. As hospitalization is a major cost factor in the management of ESRD patients, there is a strong case for face validity of the SRR measure. We used Pearson correlation coefficients to examine the relationship between the SRR and other facility-level practice patterns.

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

The measure is positively correlated with the one-year Standardized Hospitalization Ratio for Admissions (r = .53, *p* < .0001), the one-year Standardized Mortality Ratio (r = .19, *p* < .0001), and catheter use (r = .11, *p* < .0001). The SRR is negatively correlated with the percentage of patients having a Urea Reduction Ratio (URR) of at least 65% (r = -.05, *p* = .001) and using a fistula (r = -.09, *p* < .0001).

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
The SRR is a measure of hospital use, comprising many causes of hospitalization. The TEP considered

devising cause-specific SRRs but recommended the use of overall SRR measures due to various reasons, including the lack of clear consensus on which causes are modifiable by the dialysis facility and concerns about gaming the system if certain conditions are identified.

The face validity of the SRR measure is also supported by its association with other known quality measures, which include both dialysis facility outcomes and practices. Higher values of SRR are associated with higher rates of hospitalization and mortality. The SRR is also correlated with other quality measures (listed above), although the correlations are small.

In general terms, many TEP members agreed with the rationale for pursuing a readmission measure in the context of dialysis facilities, and that such a measure could help to promote shared accountability and continuity of care as dialysis patients are discharged from acute care hospitals. There were, however, two general points regarding validity that were raised and emphasized by TEP members in discussion, both at the in-person meeting and subsequently. First, several TEP members felt that it was important that the measure be adjusted for physician(s) also providing care for the patient. Second, some TEP members felt that the denominator based on the number of discharges was inappropriate and that the measure should make reference to the total number of patients under care at the facility. For the former point, it is CMS' view that dialysis facilities should be encouraged to coordinate with the nephrologists and other physicians with whom they work to reduce readmissions. We note that it is difficult to determine a unique physician associated with a discharge that could be used for adjustment, and in many cases, patients are being attended to by several physicians. It was also noted that adjustment for physician, even if possible, would mean that this measure did not harmonize in an important way with other ESRD (and general health care) measures approved by NQF and in use. It was therefore decided not to attempt any adjustment of this sort at the present time. The latter concern recognizes difficulties that arise with a random denominator. For example, a facility with a very low overall hospital utilization may, nonetheless have a high rate of readmissions. The interpretation and use of the readmission measure on its own could therefore be misleading. This issue is also discussed in the material on pairing of measures (see De.4 in the Readmission Measure Specifications), where it is noted that the Standardized Hospitalization Ratio (SHR) and the SRR should be considered together. The SHR measure appropriately reflects the level of hospital usage among patients treated by the facility with the number of patients at the facility as the reference. The SRR, on the other hand, is looking specifically at the readmission process and provides additional insight into facility outcomes, an insight that might often help to promote shared accountability between hospitals and dialysis facilities. Furthermore the empirical correlation between SHR and SRR is about 0.5, reflecting that both measures are somewhat related but not to the extent of redundancy.

2b3. EXCLUSIONS ANALYSIS
NA □ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
In the process of developing the measure of 30-day unplanned readmissions in dialysis facilities, we exclude planned readmissions from the numerator (*n* = 49,639). For details on how we determined a readmission's status as planned, please see the Appendix.

We further exclude the following hospital discharges from the denominator:
1. Not a live discharge

2. Result in a patient dying within 30 days with no readmission
3. Are against medical advice
4. Include a primary diagnosis for cancer, mental health or rehabilitation
5. Are from a PPS-exempt cancer hospital
6. Result in a transfer to another hospital on the same day
7. Occur after a patient's 12th admission in the calendar year

The numerator exclusion and first six denominator exclusions are aligned with CMS' Hospital-Wide All-Cause readmission measure. We additionally excluded discharge records following a patient's 12th admission in response to concerns from some members of the TEP held in May 2012 for this measure. Specifically, it was felt that frequently hospitalized patients would unfairly penalize smaller facilities by inflating their facility's SRR. This concern is relevant in the context of the measure's potential applications, which are to identify poor-performing facilities for quality improvement purposes. In the context of dialysis facilities, 2.8% of discharges were followed by a death within 30 days with no readmission (corresponding to exclusion #2 above). This measure concentrates on readmissions, but a complementary measure reflecting mortality within 30 days of discharge might be considered.

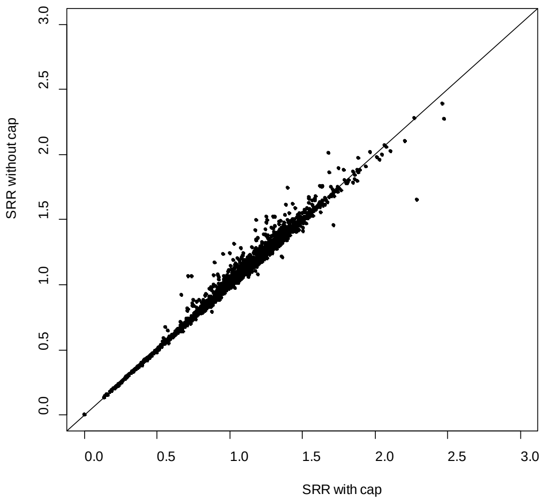
In addition, the first and sixth exclusion criteria cannot result in a readmission and so are not relevant to a readmission statistic.

We determined the cut point (cap) for admissions by examining the distribution of the number of readmissions per patient. We compared SRRs with and without the admission cap to determine the extent to which the measure changed with the exclusion.

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*)
The number and percentage of excluded discharges are as follows:
1. Not a live discharge (*n* = 31,593; 4.4%)
2. Result in a patient dying within 30 days with no readmission (*n* = 20,499; 2.8%)
3. Are against medical advice (*n* = 9,728; 1.3%)
4. Include a primary diagnosis for cancer, mental health or rehabilitation (*n* = 21,413; 3.0%)
5. Are from a PPS-exempt cancer hospital (*n* = 229; 0.03%)
6. Result in a transfer to another hospital on the same day (*n* = 21,818; 3.0%)
7. Occur after a patient's 12th admission in the calendar year (*n* = 5,155; 0.7%)

The Hospital-Wide All-Cause Readmission measure was a starting point for this measure and specified the first six exclusions. Regarding the admission-cap exclusion, we found that fewer than 1% of discharges were excluded based on this cap (0.5% of patients had more than 12 admissions in the year). As shown in Figure 1, we compared each facility's SRR with and without discharges following a patient's 12th admission in the year and found the two measures to be highly correlated (overall Pearson correlation coefficient [r] = 0.99).

Figure 1. Correlation between SRR with admission-cap and SRR without admission cap (2009).



Overall Correlation=0.99

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*)
Based on the findings presented in Figure 1, we concluded that incorporating this exclusion—which has face validity and meets the intention of the TEP—is appropriate and supported by the high degree of correlation between the measure with and without this exclusion applied.

____

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

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

☐ **No risk adjustment or stratification**
☒ **Statistical risk model with 8 categories of risk factors** *(see Table 1 in Section 2b4.4.)*
☐ **Stratification by risk categories**
☐ **Other**

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*)
The list of covariates considered was based on CMS' Hospital-Wide All-Cause Readmission Rate (HWR; NQF #1789) and CMS' Standardized Hospitalization Ratio (SHR; NQF #1463), all of which were statistically verified by the measure developer.[1] The HWR and SHR adjusted for patient comorbidities

measured at different points in time (prevalent and at ESRD incidence, respectively). Based on TEP input, we chose as a starting point the HWR comorbidity adjustments which are defined using claims data and can capture current comorbidities. There are concerns about the use of current comorbidities as adjustments in the SHR because they may reflect results of poor treatment and so lie in the causal path leading to hospitalization. These concerns are less salient when considering readmission since, whatever the cause of hospitalization, effective treatment and coordination to avoid readmission is important. In addition, we included length of the index hospitalization and severity of the index diagnosis as additional adjustments.

The risk adjustment is based on a two-stage logistic model. The adjustment is made for patient age, sex, diabetes, duration of ESRD, BMI at incidence, prior-year comorbidities, length of hospital stay and presence of a high-risk diagnosis at discharge. In the first stage of this model, both dialysis facilities and hospitals are represented as random effects, and regression adjustments are made for the set of patient-level characteristics listed above. From this first stage, we obtain the estimated standard deviation of the random effects of hospitals.

The second stage of the model is a mixed-effects model, in which facilities are fixed effects and hospitals are modeled as random effects, with the standard deviation specified as equal to its estimate from the first stage. The expected number of readmissions for each facility is estimated as the summation of the probabilities of readmission for the discharges of all patients in this facility, assuming the national average or norm for facility effect. This model accounts for a given facility's case mix using the same set of patient-level characteristics as those in the first stage.

Relevant references are below[2,3]; we conducted all analyses in R and SAS. The analyses presented here are based on ICD-9 codes; a crosswalk of ICD-9 to ICD-10 codes is presented in the Appendix.

1. Horwitz L, Partovian C, Lin Z, et al. "Hospital-wide all-cause risk-standardized readmission measure: Measure methodology report." Technical paper submitted to the Centers for Medicare and Medicaid Services. September 27, 2011. Available at http://www.naph.org/Unpublished-Documents/Hospital-Wide-All-Condition-30-Day-Risk-Standardized-Readmission-Measure.aspx. Accessed December 6, 2012.
2. He K, Kalbfleisch JD, Li Y, Li Y. "Evaluating readmission rates in dialysis facilities with or without adjustment for hospital effects." Unpublished manuscript. 2012.
3. Diggle PJ, Heagerty P, Liang KY, Zeger SL. *Analysis of Longitudinal Data (2nd ed)*. Oxford University Press; Oxford. 2002.

2b4.4. What were the statistical results of the analyses used to select risk factors?
As described above, all risk factors included in the model have face validity, and all but four—being respirator-dependent, experiencing a hip fracture/dislocation or having rheumatoid arthritis at some point in the year leading up to hospitalization, and being within 2 years of ESRD incidence—are also significantly predictive of readmission (Table 2). As the ROC curve demonstrates, the model's accuracy is fair (Figure 2); c-statistic = 0.6359.

Table 2. Covariates Included in the SRR Model

Risk Factor	Beta	SE	*p*
Age (y)			
<25	0.33	0.03	<.0001
25–45	0.18	0.01	<.0001
45–60 (ref)	—	—	—

Risk Factor	Beta	SE	*p*
60–75	-0.03	0.01	<.0001
>75	0.06	0.01	<.0001
BMI			
Underweight	0.08	0.01	<.0001
Normal Weight (ref)	—	—	—
Overweight	-0.05	0.01	<.0001
Obese	-0.12	0.01	<.0001
Cause of ESRD: Diabetes	0.05	0.01	<.0001
Comorbidity (past year)			
Amputation status	0.06	0.01	<.0001
COPD	0.22	0.01	<.0001
Cardiorespiratory failure/shock	0.23	0.01	<.0001
Coagulation defects & other specified hematological disorders	0.13	0.01	<.0001
Drug and alcohol disorders	0.32	0.02	<.0001
End-Stage Liver Disease	0.27	0.02	<.0001
Fibrosis of lung or other chronic lung disorders	0.04	0.02	0.01
Hemiplegia, paraplegia, paralysis	0.08	0.01	<.0001
Hip fracture/dislocation	0.01	0.02	0.17
Major organ transplants (excl. kidney)	-0.04	0.03	0.04
Metastatic cancer/acute leukemia	0.29	0.04	<.0001
Other hematological disorders	0.18	0.02	<.0001
Other infectious disease & pneumonias	0.15	0.01	<.0001
Other major cancers	0.02	0.01	0.04
Pancreatic disease	0.21	0.01	<.0001
Psychiatric comorbidity	0.19	0.01	<.0001
Respirator dependence/tracheostomy status	-0.03	0.04	0.11
Rheumatoid arthritis & inflammatory connective tissue disease	0.02	0.02	0.06
Seizure disorders & convulsions	0.10	0.01	<.0001
Septicemia/shock	0.13	0.01	<.0001
Severe cancer	0.15	0.02	<.0001
Severe infection	0.06	0.02	0.0002
Ulcers	0.10	0.01	<.0001
Length of Index Hospitalization (days)			
Quartile 1 (ref)	—	—	—
Quartile 2	0.12	0.01	<.0001
Quartile 3	0.23	0.01	<.0001
Quartile 4	0.44	0.01	<.0001
Presence of high-risk diagnosis at index discharge	0.49	0.03	<.0001
Sex: Female	0.06	0.01	<.0001
Time on ESRD (y)			
<1 (ref)	—	—	—
1–2	0.00	0.01	0.25
2–3	-0.32	0.01	<.0001
3–6	-0.35	0.01	<.0001
>6	-0.38	0.01	<.0001

Note. Discharge diagnoses that were relatively rare but led to a 30-day unplanned readmission in at least 40% of cases.

Figure 2. ROC curve for SRR model (c-statistic = 0.6359).

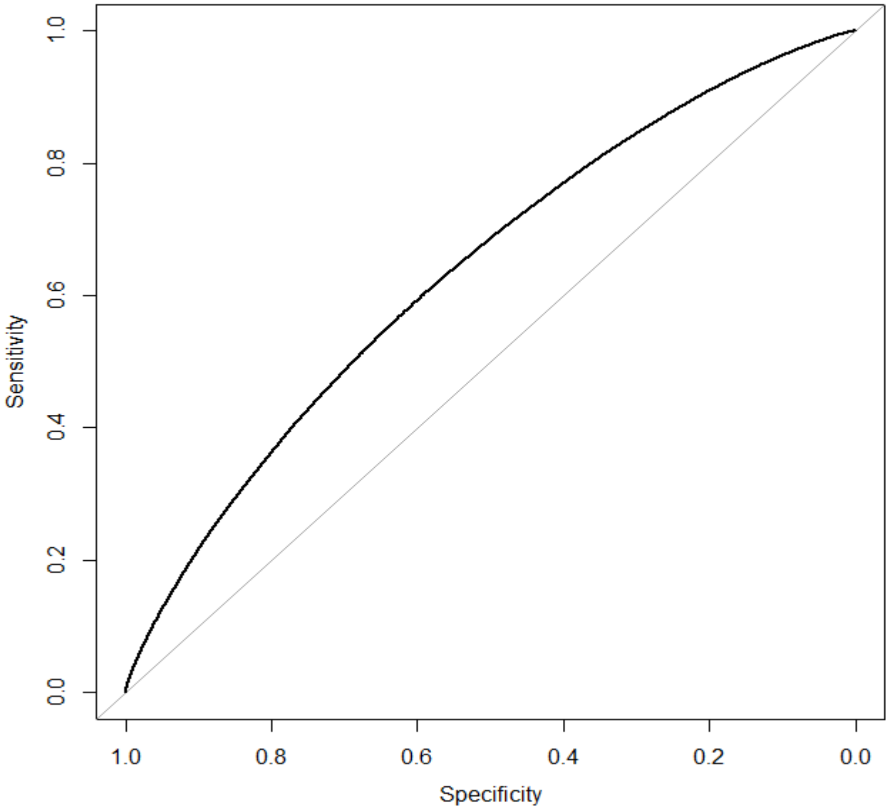
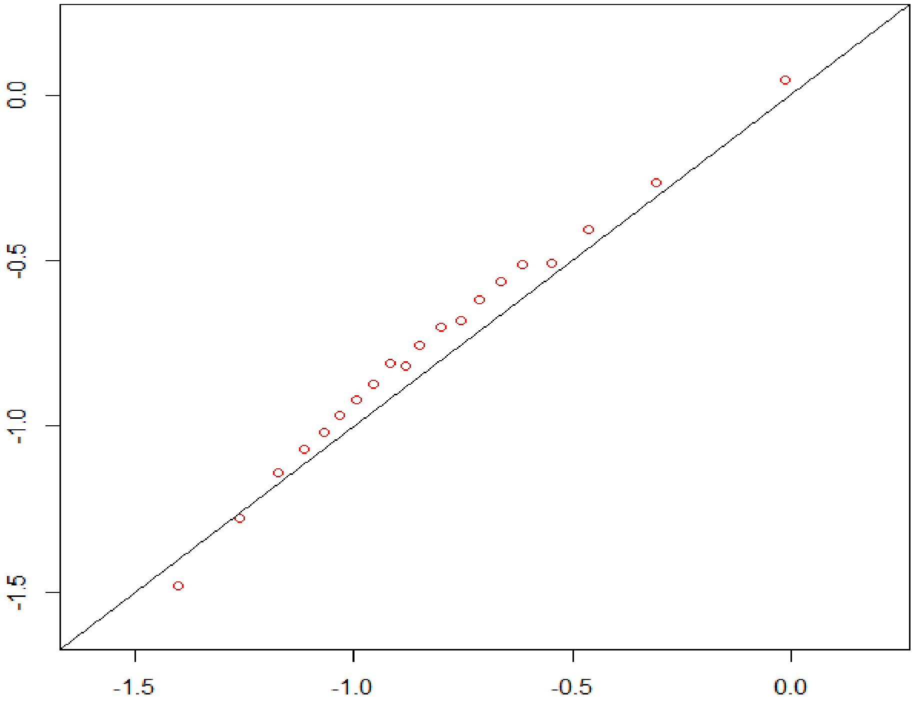


Figure 3. A plot of the logit of the observed proportion of admissions against the logit of model estimated probabilities to assess overall model fit.



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*)
The model's fit is demonstrated in Figure 2, which compares the observed rates with the model-based predictions. We bin all observations into 20 groups based on their model-based predicted values and compute the observed readmission proportion for each group. We then apply the logit transformation to each group's observed readmission proportion and plot it against the same group's average linear prediction; see the dots for all 20 groups in the plot. The 45-degree line would represent a perfect match between the observed values and the model-based predictions. In general, the closer the observed values are to this line the better the model fit. As the figure shows, the observed values are spaced fairly equally and lie very close to the 45-degree line, indicating a good fit.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
As the ROC curve demonstrates, the model's accuracy is fair (Figure 2 above); c-statistic = 0.6359.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
The Hosmer-Lemeshow test statistic based on deciles of risk is 60.9 with P-value<0.0001 (df=8). In very large samples such as this even relatively small departures from the model (such as those illustrated in Figure 2) will lead to significant results. As noted earlier, Figure 2 illustrates that the model provides an overall good fit to the data.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
The model's fit is demonstrated in Figure 3, which compares the observed rates with the model-based predictions. We bin all observations into 20 groups based on their model-based predicted values and compute the observed readmission proportion for each group. We then apply the logit transformation to each group's observed readmission proportion and plot it against the same group's average linear prediction; see the dots for all 20 groups in the plot. The 45-degree line would represent a perfect match between the observed values and the model-based predictions. In general, the closer the observed values are to this line the better the model fit.

2b4.9. Results of Risk Stratification Analysis:
An assessment of the risk analysis is given in Figure 3 above.

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*)
As Figure 3 shows, the observed values are spaced fairly equally and lie very close to the 45-degree line. This means that the model fit is good and therefore adequately adjusts for patient characteristics (case mix).

***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*)

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)*
To test the null hypothesis that the SRR for a given facility is statistically different from the national average, we use a simulation method to calculate the nominal p-value as the probability that the observed number of readmissions should be at least as extreme as that expected. This calculation is based on the supposition that, having adjusted for case mix, this facility has a true readmission rate corresponding to the average facility. Our approach captures the most important aspects of the variability in the SRR. It also avoids difficulties with more traditional methods based on estimates and standard errors. Methods are described in detail in He et al. (2013).

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*)
To address the problem of simultaneously monitoring a large number of facilities and to take account of the intrinsic unexplained variation among facilities, we used the approach described in Kalbfleisch and Wolfe (2013). This method is based on the empirical null as described in Efron (2004, 2007). The p-value for each facility is converted to a Z-score, stratified into three groups based on numbers of discharges within each facility. The empirical null corresponds to a normal curve that is fitted to the center of each Z-score histograms using a robust M-estimation method. The standard deviation of empirical null distribution is then used for a reference distribution (with mean 0) to identify outlier facilities. This method aims to separate underlying intrinsic variation in facility outcomes from variation that might be attributed to poor (or excellent) care.

The flagging rates presented in Table 3 are based on flagging those facilities in the upper tail (area=5%) of the empirical null distribution in each stratum. (The empirical null p-value is 5% or less.)

Table 3. Facilities Identified as Performing Worse than Expected for 30-Day Readmission Rate

Facility Size (No. of Patients)	**No. of Facilities**	**SRR: Worse than Expected**
Small (<=70)	1732	89 (5.14%)
Medium (71–121)	1784	126 (7.06%)
Large (>121)	1757	147 (8.37%)
Total	**5273**	**362 (6.87%)**

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?*)
Without empirical null methods, a large number of facilities will be flagged, including many larger

facilities with a relatively small difference between the rates of readmission. In contrast, the methods based on the empirical null make appropriate adjustments for overdispersion. Using this method, facilities are flagged if they have outcomes that are extreme when compared to the variation in outcomes for other facilities of a similar size.

____

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

<u>**Note**</u>*: 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*)

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*)