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

**Measure Number** (*if previously endorsed*)**:** 0291

**Measure Title**: Emergency Transfer Communication Measure

**Date of Submission**: Fall 2019

**Type of Measure:**

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

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b3.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b4.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

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

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

The Federal Office of Rural Health Policy (FORHP) has set up a reporting process for the Emergency Department Transfer Communication Measures (EDTC) Critical Access Hospitals (CAH): each CAH provides data to the State Flex Office or the State office of Rural Health or Primary Health, which is then compiled into an Excel template supplied by FORHP. The raw Excel data file from each state is submitted to FORHP, which subsequently submits that data to Telligen. Telligen then generates state and hospital reports, which are distributed back to State Flex Offices via FORHP Project Officers. State Flex Coordinators and critical access hospitals (CAHs) utilize the EDTC reports to implement quality improvement initiatives.

Data elements related to the new (revised 2019 and submitted here for re-endorsement) specifications (8 elements) are included in this data. Data from the remaining previous specifications of 21 elements are not included in the data for these analysis.

The data used to generate Cohen’s Kappa to test the reliability of element level measurement was from a June 2019 pilot test of the 2019 specifications from 34 CAHs and Stratis Health.

IRR data from previous use is included here. The IRR activities includes previous versions of specifications. This from 2009 field test data includes the current data elements.

The Telligen database is also the source of the data on the measures used for comparison in the validity section of this document and the reliability testing of the composite measure.

**1.3. What are the dates of the data used in testing**? Quarter 4 2017, June 2019

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

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

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

No information is available about the source of the data, i.e., whether it is abstracted from paper or electronic records. The data is submitted to the central data base in an excel format. CAH hospitals are not required to identify if the abstraction is from paper or electronic records.

All records that were available were used. The analysis was done on a sample of submissions but using all of the submissions available for these measures in the time frame.

For Reliability testing 1348 CAH hospitals’ data was used in the Beta-Binomial calculations. This included 190650 patient charts that are aggregated to the hospital level. Only the 8 elements from the new specifications are included in these analysis.

For IRR reliability testing 34 hospitals with approximately 83 patient chart records were used. Twenty-one (of 104) charts were excluded because the submission criteria were not followed. Raw agreement between the expert reviewer and hospital reviewers is also presented.

For the first quarter of 2009 data, 197 records were abstracted at 23 hospitals in three states. In 134 or 68% of those records the hospital abstractors findings agreed 100% with the QIO staff abstraction. In the second quarter, 165 charts were abstracted at 19 hospitals. 136, 82.4%, of those records the hospitals abstraction findings agreed 100% with the QIO staff abstraction. The number of inconsistencies in abstraction decreased from 74 to 29 from the first quarter to the second quarter.

For Validity testing 110 entities were included in the Fibrinolytics measure, 200 entities were included in the Aspirin measure and 907 entities were included in the Immunization measure for analysis. The number was based on the number of hospitals who had submitted data for the comparison measures.

|  |  |
| --- | --- |
| Number of entities | EDTC |
| Immunization | 907 |
| Administration of Fibrinolytics | 110 |
| Administration of Aspirin | 200 |

**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*)

All patients who were transferred from the emergency department to another facility that provided onsite care provided by trained medical professionals are included. Diagnosis, age race, gender, etc., were not considered in the inclusion criteria.

For reliability Beta Binomial calculations approximately 190650 patients were included in the data.

For reliability/IRR calculations approximately 83 patients were included in the data.

For reliability/IRR calculations using previously reported 2009 data 197 records were abstracted at 23 hospitals in three states.

For Validity testing – the EDTC sample compared to the comparison measures IMM had 35777 patients, OP4 had 7872 patients, and OP2 had 4527 patients.

|  |  |
| --- | --- |
| Number of patient charts | EDTC elements |
| Immunization | 35777 |
| Administration of Fibrinolytics | 4527 |
| Administration of Aspririn | 7872 |

**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**.

For reliability Beta Binomial calculations approximately 190650 patients were included from Telligen data.

For reliability/IRR calculations approximately 83 patients were included from the June 2019 Pilot IRR data.

For reliability/IRR calculations using previously reported 2009 data 197 records were abstracted at 23 hospitals in three states.

For Validity testing – the EDTC sample compared to the comparison measures IMM had 35777 patients, OP4 had 7872 patients, and OP2 had 4527 patients.

|  |  |
| --- | --- |
| Number of patient charts | EDTC elements |
| Immunization | 35777 |
| Administration of Fibrinolytics | 4527 |
| Administration of Aspirin | 7872 |

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

We have no reason to suggest that a hospital discharge or transfer processes would be influenced by social factors of the individual patients.

No patient level data was used.

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

We conducted reliability testing at the performance score level using a Beta-Binomial model and on the individual element level using Kappa score from Inter Rater Reliability (IRR).

All analysis and data cleaning were completed using SAS 9.4 in SAS Studio 3.8.

Performance score level analysis was estimated using Adams methods as in <https://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR653.pdf>.

The macro for calculating the alpha and beta values for our beta-binomial distribution was developed by:

Ian Wakeling - Qi Statistics (email: [ian@qistatistics.co.uk](mailto:ian@qistatistics.co.uk)) Web site: [www.qistatistics.co.uk](http://www.qistatistics.co.uk)

Reference: <http://www.qistats.co.uk/BetaBinomial.html>

In this report, we estimate reliability with a beta-binomial model. The beta-binomial is a natural model for estimating the reliability of simple pass/fail rate measures. There are also computational advantages to using the beta-binomial model, which is based on the beta distribution for the “true” hospital scores. The beta distribution is a very flexible distribution on the interval from 0 to 1. The beta-binomial model assumes the hospital’s score is a binomial random variable conditional on the hospital’s true value that comes from the beta distribution.

This method underscores that reliability is not just a property of a measure set but also depends what population is used to estimate the reliability. Whether a set of measures is useful for profiling hospitals depends on how different the hospitals are from one another. Measures that may be useful in one group of hospitals may not be useful in another group with little hospital to hospital variation. Similarly, as the hospitals under study increase their performance, the reliability may decrease if the hospital to hospital variance decreases over time. This is especially true as measures hit the upper limits of their ranges.\* (\*The Reliability of Provider Profiling A Tutorial John L. Adams Prepared for the National Committee for Quality Assurance HEALTH)

IRR Methods were used to assess the reliability at the data element level.

IRR was calculated using SAS.

We reviewed 8 elements for 83 hospital records. They were first reviewed by the hospital and then reviewed by an expert. To evaluate the agreement between the hospital and the expert, we used Cohen’s Kappa coefficients as an estimation of agreement. Cohen’s Kappa is widely known as an appropriate measure of agreement when looking at binary (pass/fail) agreement between two reviewers. A Kappa coefficient of 0 indicates that there was no negative or positive correlation between the two raters. Alternatively, a value of negative 1 indicates complete disagreement while positive 1 indicates complete agreement. There is no official rating of agreement in terms of strength but Kappa coefficients above 0.75 can be seen as substantial positive agreement1.

The raw agreement between the expert reviewer and the hospital reviewers showed high level of agreement with opportunity for clarification. Please note these are with the new specification revisions. Additional training will be provided Q4 2019.

Element\_1 Element\_2 Element\_3 Element\_4 Element\_5 Element\_6 Element\_7 Element\_8

73.1% 69.2% 73.1% 72.1% 70.2% 81.7% 75.0% 75.0%

For the first quarter of 2009 data, 197 records were abstracted at 23 hospitals in three states. In 134 or 68% of those records the hospital abstractors findings agreed 100% with the QIO staff abstraction. In the second quarter, 165 charts were abstracted at 19 hospitals. 136, 82.4%, of those records the hospitals abstraction findings agreed 100% with the QIO staff abstraction. The number of inconsistencies in abstraction decreased from 74 to 29 from the first quarter to the second quarter.

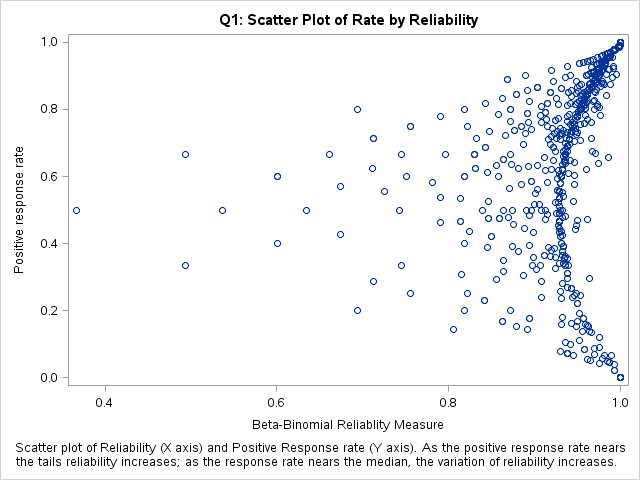
**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

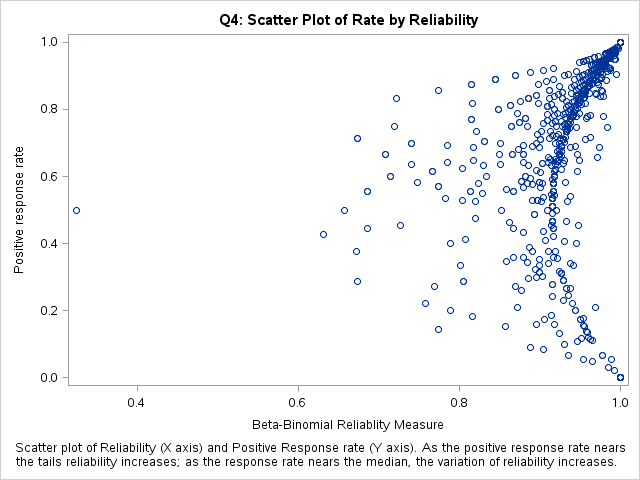
Performance measure score reliability was tested using a Beta-Binomial distribution.

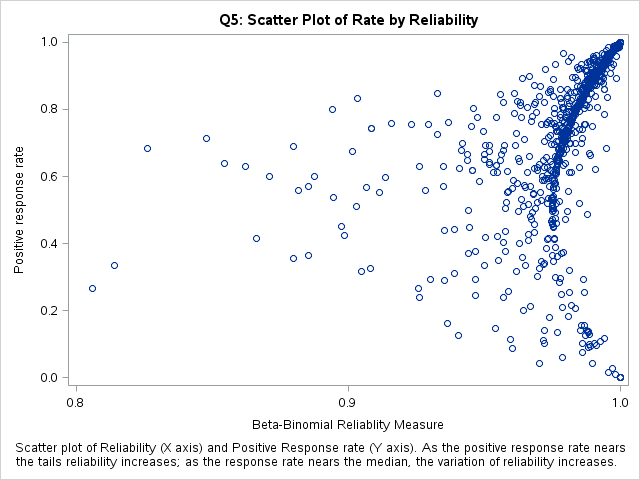
Rate and Reliability Distributions – (n: 1,185)

Reliability from Beta-Binomial Distribution (alpha = 1.4651; beta = 0.4188)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Mean | Std Dev | Min. | 10th % | 25th % | 50th % | 75th % | 90th % | 95th % | Max. |
| Q1 Score (range: 0-1) | 0.786 | 0.253 | 0 | 0.407 | 0.689 | 0.878 | 0.978 | 1.0 | 1.0 | 1.0 |
| Q1 –Reliability (range: 0-1) | 0.980 | 0.060 | 0.324 | 0.890 | 0.926 | 0.965 | 1.0 | 1.0 | 1.0 | 1.0 |







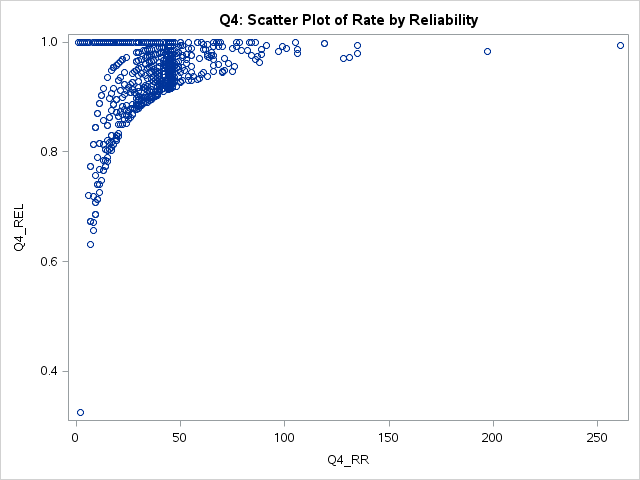
The raw agreement between the expert reviewer and the hospital reviewers showed high level of agreement with opportunity for clarification. Please note these are with the new specification revisions. Additional training will be provided Q4 2019.

Element\_1 Element\_2 Element\_3 Element\_4 Element\_5 Element\_6 Element\_7 Element\_8

73.1% 69.2% 73.1% 72.1% 70.2% 81.7% 75.0% 75.0%

**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?*)

* Interpretation: Based on the observed average reliability of 0.95 in both quarters 1 and 4, and the patterns of variation observed on the visualizations above, we can say that these measures were reliable. A key observation is that, for hospitals with the highest scores, the reliability of the measure increases, the same can be said for the hospitals with the lowest scores. Most of the variation in reliability occurs in the hospitals with a success rate between 0.2 & 0.6, though in both quarters 75% of the hospitals had a success rate above 65%. More analysis should be done into the drivers of reliability and inter hospital variation. As a visual analysis, below is a plot of reliability by number of trials for each hospital in quarter 4.
* Scatter plot of Reliability (Y) axis and number of responses (X) axis: As the response rates increase reliability increases.



* Scatter plot of Reliability (Y) axis and number of responses (X) axis: As the response rates increase reliability increases.

IRR results

There are several operational definitions of "inter-rater reliability", reflecting different viewpoints about what is a reliable agreement between raters.[[1]](https://en.wikipedia.org/wiki/Inter-rater_reliability#cite_note-1) The operational definitions of agreement: Reliable raters agree with the "official" rating of a performance.

We tested the agreement between the independent hospital based raters with the ‘official’ rating (Stratis Health, Robyn Carlson) of abstraction.

The raw agreement between the expert reviewer and the hospital reviewers showed high level of agreement with opportunity for clarification. Please note these are with the new specification revisions. Additional training will be provided Q4 2019.

Element\_1 Element\_2 Element\_3 Element\_4 Element\_5 Element\_6 Element\_7 Element\_8

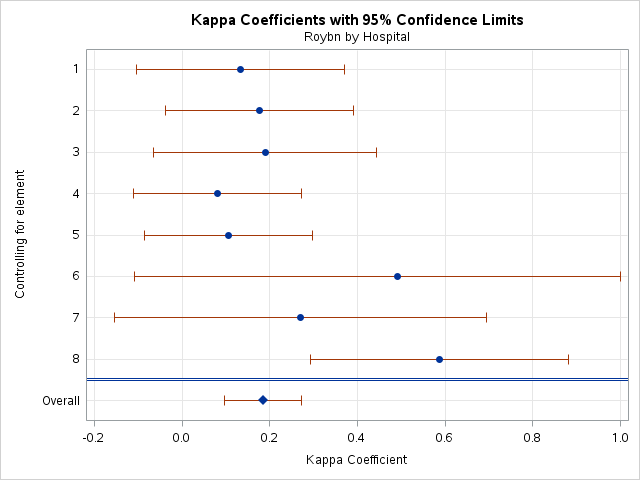
73.1% 69.2% 73.1% 72.1% 70.2% 81.7% 75.0% 75.0%

Overall, the hospital and expert had slight to fair crude non-zero agreement (Kappa = 0.224; 95% CI: [0.122 - 0.327]). The individual elements had a range of Kappa values, element 4 had the lowest agreement (Kappa=0.08) while element 5 had the highest (Kappa=0.59). The Element adjusted overall agreement within our sample was slightly lower than our observed crude value (Kappa = 0.185; 95% CI: [0.097 - 0.273]). Figure 1 below shows the element specific and overall Kappa coefficients along with a. 95% confidence interval. An overall chi-square test indicated that there was statistically significant difference between the individual element’s agreement values (p=0.171).

Our analysis found that both adjusted/unadjusted agreement between the hospital and expert raters was slight to fair. There was also no evidence that a single element of the 8 was responsible for this result as the kappa coefficients were homogeneous, though the elements with the lowest agreement were 4, 5, & 1 respectively.

Previous IRR analysis showed good agreement. The specifications with respect to the measures have changed little. The difference between the two may be due to the availability of resources. The 2019 pilot of specifications purposely did not provide support during the data collection period to test the use of the manual. Clearly the additional on call support is necessary for ongoing reliability.

For the first quarter of 2009 data, 197 records were abstracted at 23 hospitals in three states. In 134 or 68% of those records the hospital abstractors findings agreed 100% with the QIO staff abstraction. In the second quarter, 165 charts were abstracted at 19 hospitals. 136, 82.4%, of those records the hospitals abstraction findings agreed 100% with the QIO staff abstraction. The number of inconsistencies in abstraction decreased from 74 to 29 from the first quarter to the second quarter.



References:

1Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. Fam med, 37(5), 360-363.

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

**2b1. VALIDITY TESTING**

**2b1.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

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

**2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Pearson correlation coefficients were calculated comparing three ED quality performance measures that are related to process of care to the EDTC measures.

Three ED measures were used to compare with groupings of the 8 EDTC elements included in the new 2019 EDTC measurement. EDTC data is not available for the each element but subsections 4, 5, and 7 represent the elements that remain in the revised measure.

2019 EDTC Elements by subsection are

* Subsection 4
  + Medications Administered in ED
  + Allergies and Reactions
  + Home Medications
* Subsection 5
  + ED Provider Note
  + Mental Status and Orientation Assessment
  + Reason for Transfer and Plan of Care
* Subsection 7
  + Tests and Procedures Performed
  + Tests and Procedures Results

The comparison measures are IMM2 (Immunization rate of Pts), OP 2 (Fibrinolytic therapy received in the ER within 30 minutes), and OP 4 (Aspirin upon ER arrival).

Analysis was done using a one-tailed model.

The sample size for IMM2 is the highest and most amenable to comparison. Sample size for OP 4 is acceptable. Because CAH hospitals rarely provide fibrinolytic care the sample sizes are limited for OP 2.

Data for this testing was from Q4, 2017

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

Pearson Correlation Coefficients were generated comparing three Emergency Department Process measures to the EDTC element combinations.

The Comparison Measures are:

IMM2 -Immunization screening for influenza vaccine status and vaccinated prior to transfer

OP2 - Fibrinolytic Therapy Received Within 30 Minutes of ED arrival

OP4 - Aspirin therapy within 24 hours before ED arrival or prior to transfer

These comparison measures were chosen because they represent 0-1 possible outcomes and because they are process measures for patients in the same setting the ED.

The EDTC data that is used in this analysis is separate into 3 groups of elements. This is done because our data is from the previous EDTC specifications. The new specifications only include the elements in these 3 subsets. The data we have cannot be separated into the 8 individual elements nor can they be aggregated into one score.

The quality constructs for these comparisons apply to the same population (Emergency Department patients) at the same care location (Emergency departments), at the same type of facilities (CAHs), from the same data sources, and in the same time frame. These quality measures are all process measures with 0-1 results.

These quality constructs expressed by these EDTC measures vary from the comparison measures because of the nature of the processes and the facility structures that support these measures. We did not expect high correlation scores but did expect positive and statistically significant correlation.

As expected the results showed positive and significant correlation between IMM, OP4 and EDTC 4, 5, 7.

Correlations – One Tailed – Q4

In the table below the Pearson Correlation coefficients for the Immunization Rate and ER Communication measures in Q4 are presented. The Immunization Rate of patients in the ER is positively and significantly correlated to all three of the Emergency Department Transfer Communication measure, at the 1% level of significance. The implication is that increased levels of communication in the ER are positively related to higher levels of immunization in older patients.

CAHs with attention to ED quality of care follow standards of care for immunizations and communication processes. Both administration of proper influenza immunization and communication of appropriate information signal an attention to future health and ongoing care for these similar populations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlations- Q4 Immunization and EDTC 4, 5, 7 | | | | | |
|  | | imm2\_ratio\_Q4 | Q4\_EDTC4 | Q4\_EDTC5 | Q4\_EDTC7 |
| imm2\_ratio\_Q4 | Pearson Correlation | 1 | .236\*\* | .255\*\* | .218\*\* |
| Sig. (1-tailed) |  | .000 | .000 | .000 |
| N | 984 | 907 | 907 | 907 |
| Q4\_EDTC4 | Pearson Correlation | .236\*\* | 1 | .971\*\* | .976\*\* |
| Sig. (1-tailed) | .000 |  | .000 | .000 |
| N | 907 | 1185 | 1185 | 1185 |
| Q4\_EDTC5 | Pearson Correlation | .255\*\* | .971\*\* | 1 | .980\*\* |
| Sig. (1-tailed) | .000 | .000 |  | .000 |
| N | 907 | 1185 | 1185 | 1185 |
| Q4\_EDTC7 | Pearson Correlation | .218\*\* | .976\*\* | .980\*\* | 1 |
| Sig. (1-tailed) | .000 | .000 | .000 |  |
| N | 907 | 1185 | 1185 | 1185 |
| \*\*. Correlation is significant at the 0.01 level (1-tailed). | | | | | |

In the table below the Pearson Correlation coefficients for Fibrinolytic therapy and ER Communication measures in Q4 are presented. While the Fibrinolytic therapy received in the ER within 30 minutes is positively correlated to all three of the Emergency Department Transfer Communication measure, the values are not significant. There is no statistical support for the expectation that increased levels of communication in the ER are positively related to rapid Fibrinolytic therapy delivery in the ER.

Few CAHs provide urgent ER based fibrinolytic treatment for chestpain/AMI patients prior to transfer, therefore this sample is very small.

CAHs with attention to ED quality of care follow standards of care for fibrinolytic and communication processes. Both administration of proper fibrinolytics and communication of appropriate information signal an attention to future health and ongoing care for these similar populations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlations – Q4 Fibrinolytic Administration and EDTC 4, 5, 7 | | | | | |
|  | | op2\_ratio\_Q4 | Q4\_EDTC4 | Q4\_EDTC5 | Q4\_EDTC7 |
| op2\_ratio\_Q4 | Pearson Correlation | 1 | .055 | .076 | .062 |
| Sig. (1-tailed) |  | .286 | .214 | .261 |
| N | 119 | 110 | 110 | 110 |
| Q4\_EDTC4 | Pearson Correlation | .055 | 1 | .971\*\* | .976\*\* |
| Sig. (1-tailed) | .286 |  | .000 | .000 |
| N | 110 | 1185 | 1185 | 1185 |
| Q4\_EDTC5 | Pearson Correlation | .076 | .971\*\* | 1 | .980\*\* |
| Sig. (1-tailed) | .214 | .000 |  | .000 |
| N | 110 | 1185 | 1185 | 1185 |
| Q4\_EDTC7 | Pearson Correlation | .062 | .976\*\* | .980\*\* | 1 |
| Sig. (1-tailed) | .261 | .000 | .000 |  |
| N | 110 | 1185 | 1185 | 1185 |
| \*\*. Correlation is significant at the 0.01 level (1-tailed). | | | | | |

In the table above the Pearson Correlation coefficients for provision of Aspirin and ER Communication measures in Q4 are presented. The Aspirin delivery rate upon ER arrival is positively and significantly correlated to all three of the Emergency Department Transfer Communication measure, at the 5% level of significance. The implication is that increased levels of communication in the ER are positively related to higher levels of Aspirin delivery upon arrival to the ER.

CAHs with attention to ED quality of care follow standards of care for Aspirin administration for patients with chest pain and communication processes. Both administration of proper Aspirin and communication of appropriate information signal an attention to future health and ongoing care for these similar populations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlations Q4 Aspirin administration and EDTC 4, 5, 7 | | | | | |
|  | | op4\_ratio\_Q4 | Q4\_EDTC4 | Q4\_EDTC5 | Q4\_EDTC7 |
| op4\_ratio\_Q4 | Pearson Correlation | 1 | .174\*\* | .165\*\* | .164\* |
| Sig. (1-tailed) |  | .007 | .010 | .010 |
| N | 215 | 200 | 200 | 200 |
| Q4\_EDTC4 | Pearson Correlation | .174\*\* | 1 | .971\*\* | .976\*\* |
| Sig. (1-tailed) | .007 |  | .000 | .000 |
| N | 200 | 1185 | 1185 | 1185 |
| Q4\_EDTC5 | Pearson Correlation | .165\*\* | .971\*\* | 1 | .980\*\* |
| Sig. (1-tailed) | .010 | .000 |  | .000 |
| N | 200 | 1185 | 1185 | 1185 |
| Q4\_EDTC7 | Pearson Correlation | .164\* | .976\*\* | .980\*\* | 1 |
| Sig. (1-tailed) | .010 | .000 | .000 |  |
| N | 200 | 1185 | 1185 | 1185 |
| \*\*. Correlation is significant at the 0.01 level (1-tailed). | | | | | |
| \*. Correlation is significant at the 0.05 level (1-tailed). | | | | | |

**2b1.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

CAHs with attention to ED quality of care follow standards of care for Influenza Immunizations screening and administration, Aspirin administration for patients with chest pain, and communication processes. Both administration of proper administration of Influenza Immunizations, Aspirin for those with chest pain, and communication of appropriate information signal an attention to future health and ongoing care for these similar populations.

The quality constructs vary due to attention to the process of care delivery vs the process of communication. We expect a weak but positive and significant relationship.

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

**2b2. EXCLUSIONS ANALYSIS**

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

All patients transferred from an ED to another skilled care providing facility are eligible for this measure.

**2b2.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

**2b2.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

**2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

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

**2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

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

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

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

**2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

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

**2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors** *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

**2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

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

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

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

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

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

**2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

**2b3.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

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

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Data from a CMS directed a special innovation project lead by Stratis Health provided data from 100 CAH in 8 states. This data provided insight into score differences between discharge destinations.

Tellegin data was used togenerate description of differences across the entities with simple descriptive statistics. A sample of 1185 CAH reported EDTC 4, 5 and 7 and are reported here.

**2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The EDTC measures have been developed and used since 2006. EDTC improvements have been supported by HRSA, FORHP, Flex program and QIOs. Over time improvements have been demonstrated overall. Hospitals that are actively improving EDTC communication have shared process support tools, hospital transfer agreements, and other communication process mechanisms.

Once the process has been restructured with training and tools EDTC hospital to hospital scores often reach 100%. Hospital to non-acute facilities communication has not improved at the same rate or to the 100% scoring.

Stratis Health lead a one-year Centers for Medicare & Medicaid (CMS) special innovation project to assist eight Medicare Quality Improvement Organizations (QIOs), in Iowa, Maine, Missouri, Nebraska, Oklahoma, West Virginia, Wisconsin, and Wyoming, to train CAHs to collect and report seven composite ED transfer communication (EDTC) measures, identify gaps and opportunities for improvement, and begin planning to improve the transfer communication process and results.

The EDTC measures were originally developed by Stratis Health and the University of Minnesota Rural Health Research Center and endorsed by the National Quality Forum in 2007. Nearly 100 critical access hospitals (CAHs) across eight states worked on using measures to evaluate communication for transitions of care during emergency department (ED) transfers. Participating CAHs abstracted medical records to collect data on the EDTC measures. CAHs submitted data through their QIOs to Stratis Health for benchmarking with other participating facilities.

Improved process measures—56% relative improvement rate

Participating CAHs increased their percentage of medical records meeting all of the EDTC data elements over the course of the project from 28.26 to 44.13 percent—for a relative improvement rate of 56 percent.

The hospitals used the results to develop and implement improvements focused on better documentation and communication processes.

Analysis of different receiving facilities, and different levels of care data transfer analysis showed that CAH performance on the measures varied with where patients were transferred.

The percentage of medical records containing all necessary patient data transferred in a timely manner was 36.79 percent for acute care hospital transfers, but only 20.19 percent for transfers to other health care facilities, such as nursing homes.

Medical Records With All EDTC Patient Data

|  |  |
| --- | --- |
| CAH Transfers To | Percentage Complete |
| Acute care hospital | 36.7 |
| Other health care facilities | 20.19 |

This data highlights an opportunity for improved transition communication from EDs to long term care facilities, by working with local nursing homes to develop standard communication and transition processes.

The EDTC measure is included in phase three of the Medicare Beneficiary Quality Improvement Project (MBQIP). Starting fall 2014, CAHs nationwide can collect and submit the EDTC measures. MBQIP is a program of the Health Resources and Services Administration (HRSA) funded Office of Rural Health Policy’s (ORHP) Medicare Rural Hospital Flexibility Program (Flex).

Data from Telligen does not identify destination therefore the significance difference between entities for acute and non-acute transfers cannot be generated.

Below are descriptive statistics of the distribution of Quarter 4 EDTC 4, EDTC 5, EDTC 7 from the Telligen data. This data does not differentiate between acute and non-acute transfers.

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
| *q4 edtc 4 %* | |
|  |  |
| Mean | 0.93 |
| Standard Error | 0.00 |
| Median | 1.00 |
| Mode | 1.00 |
| Standard Deviation | 0.14 |
| Sample Variance | 0.02 |
| Kurtosis | 11.91 |
| Skewness | -3.10 |
| Range | 1.00 |
| Minimum | 0.00 |
| Maximum | 1.00 |
| Sum | 1096.39 |
| Count | 1185.00 |
|  |  |
|  |  |
| *q4 edtc 5 %* | |
|  |  |
| Mean | 0.93 |
| Standard Error | 0.00 |
| Median | 1.00 |
| Mode | 1.00 |
| Standard Deviation | 0.14 |
| Sample Variance | 0.02 |
| Kurtosis | 13.71 |
| Skewness | -3.38 |
| Range | 1.00 |
| Minimum | 0.00 |
| Maximum | 1.00 |
| Sum | 1107.68 |
| Count | 1185.00 |
|  |  |
| *q4 edtc 7 %* | |
|  |  |
| Mean | 0.96 |
| Standard Error | 0.00 |
| Median | 1.00 |
| Mode | 1.00 |
| Standard Deviation | 0.09 |
| Sample Variance | 0.01 |
| Kurtosis | 33.26 |
| Skewness | -4.73 |
| Range | 1.00 |
| Minimum | 0.00 |
| Maximum | 1.00 |
| Sum | 1142.43 |
| Count | 1185.00 |

Scatter plot Distribution of Measure Scores for EDTC subsection 4 (Y) axis and Observations (X) axis. Display shows a mean of .93 and a SD of 0.14

Scatter plot Distribution of Measure subsection 5 Scores (Y) axis and observations (X) axis. Display shows a mean of .93 and a SD of 0.14

Scatter plot Distribution of Measure Subsection 7 Scores (Y) axis and observations (X) axis. Display shows a mean of .96 and a SD of 0.09.

**2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

The data include transfers to all facilities that provide clinically trained staff. CAH to tertiary hospital communication has improved over time. CAH transfers to non-acute facilities such as nursing homes, assisted living, detox centers still needs improvement. That data assessment is not differentiated in this data set. Changes in submission expectations are being considered to help identify these improvement opportunities.

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

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

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

**Note***: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

**2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

**2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

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

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

**2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

The EDTC measure looks for completeness of data elements in a communication between two facilities. Its purpose is to identify the information that is missing from or not identified and/or included in the communication.

Because CAHs by definition limit stays to 72 hours and have a structure/process developed with supporting hospitals to transfer patients with higher acuity or special care needs beyond what is available in the CAH; this measure attempts to measure one of CAHs’ key roles, that of a communicator of initial information to promote continuity of care and decrease redundant testing.

We examine the charts from the sending hospital as we evaluate their performance as a transferring hospital. To assess the sensitivity or specificity of missing data abstraction we would need to look at the receiving facilities records to identify cases where records were identified as sent but were not sent; or records that were identified as not sent but were in fact sent. We do not have access to the data at the receiving hospital. Without access to the receiving hospital data we are unable identify missing data nor to estimate the errors related to that.

The measure itself looks at data that might not be sent. Analysis of potential data abstraction errors was presented in the IRR Kappa statistics.

**2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

**2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)