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

**Measure Number** (*if previously endorsed*)**:** 0576

**Measure Title**: Follow-Up After Hospitalization for Mental Illness

**Date of Submission**: 8/3/2020

**Type of Measure:**

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

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

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

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

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

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

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

**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*).

**2020 Submission:**

N/A

**2016 Submission:**

N/A

**1.3. What are the dates of the data used in testing**? Click here to enter date range

**2020 Submission**

Testing of measure score reliability and validity was performed using data from calendar year 2018.

**2016 Submission:**

2009-2011

2014-2016

**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*)

**2020 Submission:**

This measure assesses the percentage of discharges for members 6 years of age and older who were hospitalized for treatment of selected mental illness or intentional self-harm diagnoses and who had a follow-up visit with a mental health provider. Two rates are reported:

1. The percentage of discharges for which the member received follow-up within 30 days after discharge.
2. The percentage of discharges for which the member received follow-up within 7 days after discharge.

Testing was completed at the health plan level which is appropriate for the level of reporting for this measure.

Measure score reliability testing and construct validity testing:  The measure score reliability was calculated from HEDIS data that included 358 Commercial health plans, 172 Medicaid plans, and 308 Medicare plans. The sample included all Commercial, Medicare and Medicaid health plans submitting data to NCQA for this HEDIS measure. The plans were geographically diverse and varied in size.

**2016 Update: MEASURE SCORE RELIABILITY TESTING**

MEASURE SCORE RELIABILITY TESTING

The measure score reliability was calculated from 2016 HEDIS data that included 368 Commercial health plans, 166 Medicaid health plans, and 301 Medicare health plans for the 7-day follow-up rate and 368 Commercial health plans, 168 Medicaid health plans, and 301 Medicare health plans for the 30-day follow-up rate. The sample included all health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

SYSTEMATIC EVALUATION OF FACE VALIDITY

The Follow-Up After Hospitalization for Mental Illness measure was tested for face validity with several panels of experts. Measurement Advisory Panels (MAP) provide the clinical and technical knowledge required to develop the measures. The Behavioral Health MAP included 12 experts in behavioral health including representation by consumers, health plans, health care providers and policy makers. NCQA’s Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 15 members. The CPM is organized and managed by NCQA, and is responsible for advising NCQA staff on the development and maintenance of performance measures. The CPM also meets with the NCQA Board of Directors to recommend measures for inclusion in HEDIS. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement. Additional HEDIS Expert Panels provide invaluable assistance by identifying methodological issues and giving feedback on new and existing measures. See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel.

**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*)

**2020 Submission**

Data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the sample. It includes number of health plans included in HEDIS data collection and the average eligible population for the measure across health plans. For this measure, the eligible population is the number of eligible discharges among plan members 6 years of age and older.

##### *7-day Follow-Up Rate*

|  |  |  |
| --- | --- | --- |
| **Product Line** | **Number of Plans** | **Mean number of eligible discharges per plan** |
| Commercial | 361 | 668 |
| Medicaid | 173 | 1946 |
| Medicare | 308 | 344 |

30-Day Follow-Up Rate

|  |  |  |
| --- | --- | --- |
| Product Line | **Number of Plans** | **Mean number of eligible discharges per plan** |
| Commercial | 358 | 665 |
| Medicaid | 172 | 1956 |
| Medicare | 308 | 344 |

**2016 Update: MEASURE SCORE RELIABILITY TESTING**

Patients included for measure score reliability testing: In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47.0 million Medicaid beneficiaries, and 17.6 million Medicare beneficiaries. Data are summarized at the health plan level and stratified by product line. Below is a description of the testing data, including number of health plans included and the mean eligible population for the measure across health plans.

##### *7-day Follow-Up Rate*

|  |  |  |
| --- | --- | --- |
| **Product Line** | **Number of Plans** | **Mean number of eligible patients per plan** |
| Commercial | 368 | 568 |
| Medicaid | 166 | 1,182 |
| Medicare | 301 | 279 |

30-Day Follow-Up Rate

|  |  |  |
| --- | --- | --- |
| Product Line | **Number of Plans** | **Mean number of eligible patients per plan** |
| Commercial | 368 | 568 |
| Medicaid | 168 | 1,169 |
| Medicare | 301 | 279 |

**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**.

**2020 Submission**

No differences in the data used for reliability and construct validity testing.

**2016 Update: MEASURE SCORE RELIABILITY TESTING**

Reliability of the measure score was tested using a beta-binomial calculation. This analysis included the entire HEDIS data for the measure (described above).

Validity was demonstrated through a systematic assessment of face validity. Per NQF instructions we have described the composition of the technical expert panel which assessed face validity in the data sample questions above.

**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.

**2020 Submission**

We did not assess data by social risk factors. Social risk factor data were not available in reported results. This measure is specified for Medicare, Medicaid and Commercial members age 6 and older. NCQA is actively engaged with partners including the CMS Office of Minority Health in identifying feasible methods to further integrate social risk factors into health plan quality measures, with a focus on stratification. This is aligned with recent recommendations from MedPAC and ASPE on optimal methods for addressing social risk in quality measurement and programs.1,2This is an NCQA wide initiative. Our intent is to implement methods to bridge data concerns in the future.

1. Medicare Payment Advisory Commission. (2020). The Medicare Advantage program: Status report. In Report to the Congress: Medicare Payment Policy (p. 397). <http://medpac.gov/docs/default-source/reports/mar20_medpac_ch13_sec.pdf>
2. Office of the Assistant Secretary for Planning and Evaluation, & U.S. Department of Health & Human Services. (2020). Second Report to Congress on Social Risk and Medicare’s Value-Based Purchasing Programs. <https://aspe.hhs.gov/social-risk-factors-and-medicares-value-basedpurchasing-programs>

**2016 Update**

Measure performance was assessed by Commercial, Medicaid, and Medicare plan types.

**2012 Submission**

The measure is not stratified to detect disparities. NCQA has participated with IOM and others in attempting to include information on disparities in measure data collection. However, at the present time, this data, at all levels (claims data, paper chart review, and electronic records), is not coded in a standard manner, and is incompletely captured. There are no consistent standards for what entity (physician, group, plan, employer) should capture and report this data. While “requiring” reporting of the data could push the field forward, it has been our position that doing so would create substantial burden with inability to use the data because of its inconsistency. At the present time, we agree with the IOM report that disparities are best considered by the use of zip code analysis which has limited applicability in most reporting situations. At the health plan level, for HEDIS health plan data collection, NCQA does have extensive data related to our use of stratification by insurance status (Medicare, Medicaid and private-commercial) and would strongly recommend this process where the data base supporting the measurement includes this information. However, we believe that the measure specifications should NOT require this since the measure is still useful where the data needed to determine disparities cannot be ascertained from the data available.

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

**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*)

**2020 Submission**

Reliability testing of performance measure score

We utilized the methodology described by John Adams (Adams, J.L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009) to calculate signal-to-noise reliability. This methodology uses the Beta-binomial model to assess how well one can confidently distinguish the performance of one reporting entity from another. Conceptually, the Beta-binomial model is the ratio of signal to noise. The signal is the proportion of the variability in measured performance that can be explained by real differences across reporting entities (plans, physicians, etc.) in performance. The Beta-binomial model is an appropriate model when estimating the reliability of simple pass/fail rate measures, such as the Follow-Up After Hospitalization for Mental Illness measure. Reliability scores range from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (i.e., noise), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance across reporting entities.

For the Follow-Up After Hospitalization for Mental Illness measure, health plans are the reporting entity. For the formulas and explanations below, we use health plans as the reporting entity.

The formula for signal-to-noise reliability is:

Signal-to-noise reliability = σ2plan-to-plan / (σ2plan-to-plan + σ2error)

Therefore, we need to estimate two variances: 1) variance between plans (σ2plan-to-plan); 2) variance within plans (σ2error).

1. Variance between plans = σ2plan-to-plan = (α β) / (α + β + 1)(α + β)2

α and β are two shape parameters of the Beta-Binomial distribution, α >0, β > 0

1. Variance within plans: σ2error = p̂(1- p̂)/n

p̂ = observed rate for the plan

n = plan-specific denominator for the observed rate (most often the number of eligible plan members; in this case, the number of eligible discharges associated with each plan)

Using Adams’ 2009 methodology, we estimated the reliability for each reporting entity, then averaged these reliability estimates across all reporting entities to produce a point estimate of signal-to-noise reliability. We label this point estimate “mean signal-to-noise reliability”. The mean signal-to-noise reliability measures how well, on average, the measure can differentiate between reporting entity performance on the measure.

Along with the point estimate of mean signal-to-noise reliability, we are also providing:

1. The standard error (SE) and 95% confidence interval (95% CI) of the mean signal-to-noise reliability for all plans and stratified by the denominator size (number of eligible members per plan). The SE and 95% CI of the mean signal-to-noise reliability provides information about the stability of reliability. The 95% CI is the mean signal-to-noise reliability ± (1.96\*SE). We also stratified the results by the denominator size using terciles of the distribution to provide additional information about the stability of reliability.
2. The distribution (minimum, 10th, 25th, 50th, 75th, 90th, maximum) of the plan-level signal-to-noise reliability estimates. Each plan’s reliability estimate is a ratio of signal to noise, as described above [ σ2plan-to-plan / (σ2plan-to-plan + σ2error)]. Variability between plans (σ2plan-to-plan) is the same for each plan, while the specific plan error (σ2error) varies. Reliability for each plan is an ordinal measure of how well one can determine where a plan lies in the distribution of reliability across all plans, with higher estimates indicating better reliability. We also stratified the results by the denominator size using terciles of the distribution to provide additional information about the distribution of plan-level signal-to-noise reliability estimates. The number of plans in each stratum and the per-plan denominators of indicators are displayed in the summary tables.

This methodology allows us to estimate the reliability for each plan and summarize the distribution of these estimates.

**2016 Update: Method for measure score reliability testing** METHOD FOR BETA-BINOMIAL RELIABILITY TESTING

The beta-binomial method (Adams, 2009) measures the proportion of total variation attributable to a health plan, which represents the *signal*. The beta-binomial model also estimates the proportion of variation attributable to measurement error for each plan, which represents *noise*. The reliability of the measure is represented as the ratio of signal to noise.

* A score of 0 indicates none of the variation (signal) is attributable to the plan
* A score of 1.0 indicates all of the variation (signal) is attributable to the plan
* A score of 0.7 or higher indicates adequate reliability to distinguish performance between two plans

PLAN-LEVEL RELIABILITY

The underlying formulas for the beta-binomial reliability can be adapted to construct a plan-specific estimate of reliability by substituting variation in the individual plan’s variation for the average plan’s variation. The reliability for some plans may be more or less than the overall reliability across plans.

Adams JL. The Reliability of Provider Profiling: A Tutorial. Santa Monica, CA: RAND Corp. TR-653-NCQA, 2009

**2012 Submission**

In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009) in work produced for the National Committee for Quality Assurance (NCQA).

The following is quoted from the tutorial which focused on provider-level assessment: “Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician’s data as well as increasing the number of measures per patient.” This approach is also relevant to health plans and other accountable entities.

The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual accountable entity variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. Adams’ approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures.

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

**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*)

**2020 Submission:**

Signal-to-Noise Reliability Assessment for the *Follow-Up After Hospitalization for Mental Illness* Measure

Table 1 shows the point estimates of mean signal-to-noise reliability using above methodology.

Table 1. Point Estimates of Mean Signal-to-Noise Reliability by Product Type, Calendar Year 2018 Data

|  |  |  |  |
| --- | --- | --- | --- |
| *Follow-Up After Hospitalization for Mental Illness* | Point estimate: Mean Signal-To-Noise Reliability | | |
| Commercial | Medicaid | Medicare |
| Follow-Up After Hospitalization For Mental Illness - 7 days | 0.884 | 0.969 | 0.900 |
| Follow-Up After Hospitalization For Mental Illness - 30 days | 0.883 | 0.967 | 0.910 |

Table 2a. Mean Signal-To-Noise Reliability, Standard Error (SE) and 95% Confidence Interval (95% CI) for the Follow-Up After Hospitalization For Mental Illness - (7 day)Measure by Terciles of the Denominator Size and for All Submissions Stratified by Plan Type, Calendar Year 2018 Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stratification | Number of Plans | Number of Eligible Discharges per Plan (min - max) | Mean Signal-To-Noise Reliability | SE | 95% CI |
| **All Commercial** | **361** | **30-7412** | **0.884** | **0.006** | **(0.872, 0.895)** |
| Tercile 1 | 120 | 30-131 | 0.783 | 0.007 | (0.769, 0.797) |
| Tercile 2 | 118 | 132-471 | 0.915 | 0.003 | (0.910, 0.920) |
| Tercile 3 | 123 | 482-7412 | 0.976 | 0.001 | (0.973, 0.978) |
| **All Medicaid** | **173** | **38-17406** | **0.969** | **0.003** | **(0.962, 0.975)** |
| Tercile 1 | 57 | 38-482 | 0.933 | 0.006 | (0.921, 0.946) |
| Tercile 2 | 57 | 512-2030 | 0.987 | 0.001 | (0.986, 0.989) |
| Tercile 3 | 59 | 2054-17406 | 0.994 | 0.000 | (0.993, 0.994) |
| **All Medicare** | **308** | **30-4224** | **0.900** | **0.005** | **(0.890, 0.909)** |
| Tercile 1 | 102 | 30-90 | 0.815 | 0.007 | (0.802, 0.828) |
| Tercile 2 | 101 | 91-270 | 0.920 | 0.003 | (0.913, 0.927) |
| Tercile 3 | 105 | 273-4224 | 0.973 | 0.002 | (0.970, 0.976) |

SE: Standard Error of the mean.

95% CI: 95% confidence interval.

Table 2b. Mean Signal-To-Noise Reliability, Standard Error (SE) and 95% Confidence Interval (95% CI) for the Follow-Up After Hospitalization For Mental Illness - (30-day)Measure by Terciles of the Denominator Size and for All Submissions Stratified by Plan Type, Calendar Year 2018 Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stratification | Number of Plans | Number of Eligible Discharges per Plan (min - max) | Mean Signal-To-Noise Reliability | SE | 95% CI |
| **All Commercial** | **361** | **30-7412** | **0.883** | **0.006** | **(0.872, 0.895)** |
| Tercile 1 | 120 | 30-131 | 0.807 | 0.007 | (0.794, 0.821) |
| Tercile 2 | 116 | 132-470 | 0.921 | 0.002 | (0.916, 0.926) |
| Tercile 3 | 122 | 471-7412 | 0.971 | 0.001 | (0.969, 0.974) |
| **All Medicaid** | **174** | **38-17406** | **0.967** | **0.004** | **(0.960, 0.975)** |
| Tercile 1 | 57 | 38-512 | 0.932 | 0.007 | (0.919, 0.945) |
| Tercile 2 | 56 | 529-2030 | 0.988 | 0.001 | (0.986, 0.989) |
| Tercile 3 | 59 | 2054-17406 | 0.994 | 0.000 | (0.993, 0.994) |
| **All Medicare** | **308** | **30-4224** | **0.910** | **0.004** | **(0.902, 0.918)** |
| Tercile 1 | 102 | 30-90 | 0.838 | 0.004 | (0.829, 0.847) |
| Tercile 2 | 101 | 91-270 | 0.933 | 0.002 | (0.929, 0.937) |
| Tercile 3 | 105 | 273-4224 | 0.974 | 0.001 | (0.971, 0.977) |

SE: Standard Error of the mean.

95% CI: 95% confidence interval.

Table 3a. Distribution of Plan-Level Signal-To-Noise Reliability for the Follow-Up after Hospitalization for Mental Illness7-daymeasure rate by Terciles of the Denominator Size and for All Submissions by Plan Type, Calendar Year 2018 Data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Distribution of Plan Estimates of Signal-to-Noise Reliability | | | | | | |
| Stratification | Number of Plans | Min | P10 | P25 | P50 | P75 | P90 | Max |
| **All Commercial** | **361** | **0.567** | **0.708** | **0.826** | **0.884** | **0.972** | **0.987** | **1.000** |
| Tercile 1 | 120 | 0.608 | 0.670 | 0.735 | 0.783 | 0.849 | 0.865 | 1.000 |
| Tercile 2 | 118 | 0.852 | 0.876 | 0.894 | 0.915 | 0.941 | 0.948 | 0.959 |
| Tercile 3 | 123 | 0.947 | 0.957 | 0.966 | 0.976 | 0.986 | 0.992 | 0.997 |
| **All Medicaid** | **173** | **0.743** | **0.916** | **0.964** | **0.969** | **0.995** | **0.997** | **0.999** |
| Tercile 1 | 57 | 0.770 | 0.867 | 0.906 | 0.933 | 0.969 | 0.973 | 0.985 |
| Tercile 2 | 57 | 0.976 | 0.979 | 0.983 | 0.987 | 0.991 | 0.993 | 0.997 |
| Tercile 3 | 59 | 0.988 | 0.991 | 0.992 | 0.994 | 0.996 | 0.997 | 0.999 |
| **All Medicare** | **308** | **0.653** | **0.771** | **0.848** | **0.900** | **0.967** | **0.987** | **1.000** |
| Tercile 1 | 102 | 0.667 | 0.721 | 0.776 | 0.815 | 0.859 | 0.902 | 1.000 |
| Tercile 2 | 101 | 0.845 | 0.869 | 0.893 | 0.920 | 0.944 | 0.964 | 0.985 |
| Tercile 3 | 105 | 0.939 | 0.952 | 0.958 | 0.973 | 0.988 | 0.993 | 0.997 |

Table 3b. Distribution of Plan-Level Signal-To-Noise Reliability for the Follow-Up after Hospitalization for Mental Illness30-daymeasure rate by Terciles of the Denominator Size and for All Submissions by Plan Type, Calendar Year 2018 Data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Distribution of Plan Estimates of Signal-to-Noise Reliability | | | | | | |
| Stratification | Number of Plans | Min | P10 | P25 | P50 | P75 | P90 | Max |
| **All Commercial** | **358** | **0.541** | **0.701** | **0.821** | **0.883** | **0.973** | **0.988** | **1.000** |
| Tercile 1 | 120 | 0.622 | 0.702 | 0.757 | 0.807 | 0.865 | 0.891 | 1.000 |
| Tercile 2 | 116 | 0.846 | 0.881 | 0.903 | 0.921 | 0.943 | 0.951 | 0.968 |
| Tercile 3 | 122 | 0.935 | 0.949 | 0.960 | 0.971 | 0.984 | 0.990 | 0.997 |
| **All Medicaid** | **172** | **0.708** | **0.908** | **0.963** | **0.967** | **0.995** | **0.997** | **0.999** |
| Tercile 1 | 57 | 0.747 | 0.860 | 0.910 | 0.932 | 0.969 | 0.975 | 0.983 |
| Tercile 2 | 56 | 0.975 | 0.979 | 0.984 | 0.988 | 0.992 | 0.994 | 0.995 |
| Tercile 3 | 59 | 0.989 | 0.991 | 0.992 | 0.994 | 0.996 | 0.997 | 0.999 |
| **All Medicare** | **308** | **0.714** | **0.788** | **0.861** | **0.910** | **0.967** | **0.989** | **0.997** |
| Tercile 1 | 102 | 0.737 | 0.779 | 0.801 | 0.838 | 0.874 | 0.893 | 0.944 |
| Tercile 2 | 101 | 0.891 | 0.905 | 0.916 | 0.933 | 0.949 | 0.958 | 0.968 |
| Tercile 3 | 105 | 0.949 | 0.955 | 0.961 | 0.974 | 0.988 | 0.992 | 0.997 |

**2016 Update: MEASURE SCORE RELIABILITY**

MEASURE LEVEL RELIABILITY

NCQA pools data reported by health plans according to product line. The mean reliability for the 7-day Rate per the beta binomial model was 0.97 for Commercial health plans, 0.96 for Medicare, and 0.99 for Medicaid. The mean reliability for the 30-day Rate was 0.96 for Commercial health plans, 0.97 for Medicare, and 0.99 for Medicaid.

Beta-Binomial Statisitc For Each Measure Rate

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rate | *Commercial* | | *Medicaid* | | *Medicare* | |
| Avg | Minimum | Avg | Minimum | Avg | Minimum |
| 7-Day Follow-Up | 1.0 | 0.7 | 1.0 | 0.8 | 1.0 | 0.7 |
| 30-Day Follow-Up | 1.0 | 0.6 | 1.0 | 0.8 | 1.0 | 0.8 |

**2012 Submission**

Rate 1. The percentage of members who received follow-up within 30 days of discharge

Commercial: 0.967434

Medicaid: 0.988749

Medicare: 0.949915

Rate 2. The percentage of members who received follow-up within 7 days of discharge.

Commercial: 0.954861

Medicaid: 0.989110

Medicare: 0.951935

**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?*)

**2020 submission:**

In general, a score of 0.7 or higher suggests the measure has adequate reliability. The results suggest the measure has high reliability and more details are discussed below.

Table 2a provides the point estimate of mean signal-to-noise reliability, its standard error, and the 95% CI for the Follow-Up After Hospitalization For Mental Illness - 7 daymeasure rate for Commercial, Medicaid and Medicare plans overall and stratified by the denominator size (distribution of the number of eligible members per plan). Over all commercial plans, the reliability estimate is 0.884, and the 95% CI is (0.872, 0.895), indicating good reliability. Stratified analyses show that reliability increase as plan size gets larger and exceeds .8 for all terciles. Over all Medicaid plans, the reliability estimate is 0.969 and the 95% CI is (0.962, 0.975), indicating very good reliability. Results from the stratified analyses show that reliability exceeds 0.9 for all terciles. Over all Medicare plans, the reliability estimate is 0.900 and the 95% CI is (0.890, 0.909), indicating very good reliability. Results from the stratified analyses show that reliability tends to increase as plan size gets larger and exceeds .8 for all terciles.

Table 2b provides the point estimate of mean signal-to-noise reliability, its standard error, and the 95% CI for the Follow-Up After Hospitalization For Mental Illness - 30 daymeasure rate for Commercial, Medicaid and Medicare plans overall and stratified by the denominator size (distribution of the number of eligible members per plan). Across all commercial plans, the reliability estimate is 0.883, and the 95% CI is (0.872, 0.895) indicating good reliability. Stratified analyses show that reliability increases as plan size gets larger and exceeds .8 for all terciles. Across all Medicaid plans, the reliability estimate is 0.967 and the 95% CI is (0.960, 0.975), indicating very good reliability. Results from the stratified analyses show that reliability exceeds 0.9 for all terciles. Across all Medicare plans, the reliability estimate is 0.910 and the 95% CI is (0.902, 0.918) indicating very good reliability. Results from the stratified analyses show that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.

Table 3a summarizes the distribution of plan-level signal-to-noise reliability estimates for the Follow-Up After Hospitalization for Mental Illness 7-day measure rate. Over all commercial plans, the estimates range from 0.567 to 0.100. The 50th percentile is 0.884, which exceeds the 0.70 threshold for reliability. For Medicaid plans, the estimates range from 0.743 to .999; the 10th percentile is 0.916, indicating very good reliability. For Medicare plans, the estimates range from 0.653 to 1.000; the 50th percentile is 0.900, which exceeds the 0.70 threshold for reliability. This table also include the distribution of plan-level signal-to-noise reliability estimates stratified by denominator size. Reliability estimates tend to be higher for plans with a larger denominator.

Table 3b summarizes the distribution of plan-level signal-to-noise reliability estimates for the Follow-Up After Hospitalization for Mental Illness 30-day measure rate. Over all commercial plans, the estimates range from 0.541 to 1.000. The 50th percentile is 0.883, which exceeds the 0.70 threshold for reliability. For Medicaid plans, the estimates range from 0.708 to 0.999; the 10th percentile is 0.908, indicating very good reliability. For Medicare plans, the estimates range from 0.714 to 0.997; the 50th percentile is 0.910, which exceeds the 0.70 threshold for reliability. This table also include the distribution of plan-level signal-to-noise reliability estimates stratified by denominator size. Reliability estimates tend to me higher for plans with a larger denominator.

**2016 submission:**

Among Commercial, Medicare, and Medicaid plans, results indicate both the 7-day and 30-day rates within this measure have a good signal to noise ratio that are well above the 0.7 threshold for adequate reliability. This data analysis suggests the measure has high reliability and can discriminate performance between accountable entities.

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

**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)*

**2020 Submission**

We assessed construct validity and face validity for this measure.

Method of testing construct validity

We tested for construct validity by exploring the following:

* Are the individual rates within the *Follow-Up After Hospitalization for Mental Illness* measure positively correlated with one another?
* Is the *Follow-Up After Hospitalization For Mental Illness*  measure positively correlated with the HEDIS *Follow-Up After Emergency Department Visit for Mental Illness* measure which assesses emergency department (ED) visits for adults and children 6 years of age and older with a diagnosis of mental illness and who received a follow-up visit for mental illness within 7- and 30-days*?*

We hypothesized that rates within the *Follow-Up After Hospitalization For Mental Illness* measure would be highly positively correlated, and that organizations that perform well on *Follow-Up After Hospitalization For Mental Illness* should perform well on the other measure, *Follow-Up After Emergency Department Visit for Mental Illness*, given that they address the same or similar populations and that they address similar activities for patients following an acute event involving mental illness.

NCQA performs Pearson correlation for construct validity using HEDIS health plan data. The test estimates the strength of linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a strong positive linear association: an increase in values of one variable is associated with increase in value of another variable. A value of 0 indicates no linear association. A value of -1 indicates a strong negative relationship in which an increase in values of the first variable is associated with a decrease in values of the second variable. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The sample size for the correlation analysis is the number of plans that reported both measures. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We adjusted our p-values to account for testing multiple correlations and used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone.

Method of assessing face validity

NCQA develops measures using a standardized process. For new measures, face validity is assessed at various steps as described below.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical measurement advisory panels (MAPs), whose members are authorities on clinical priorities for measurement, participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness, and feasibility. This information is gathered into a work-up format, which is vetted by the MAPs, including the Behavioral Health Measurement Advisory Panel (BHMAP), Geriatric Measurement Advisory Panel (GMAP), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. At this step, face validity is systematically determined by the CPM, which uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA about proposed new measures. Public comment offers an opportunity to assess the validity, feasibility, importance and other attributes of a measure from a wider audience. For this measure, a majority of public comment respondents supported the measure. NCQA MAPs and the technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM.  Face validity is then again systematically assessed by the CPM. The CPM reviews all comments before making a final decision and votes to recommend approval of new measures for HEDIS. NCQA’s Board of Directors then approves new measures.

**2016 Update**

Method of Assessing Face Validity

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure “Desirable”? The work-up is vetted by NCQA’s Measurement Advisory Panels (MAPs) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.

NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA’s State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA’s experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure’s performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA’s Policy Clarification Support portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year’s data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year’s HEDIS Volume 2.

#### ICD-10 CONVERSION

The below steps describe our methods to convert this measure to ICD-10 in order to develop a new code set fully consistent with the intent of the measure.

1. NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use General Equivalence Mapping (GEM) to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
2. NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.
3. NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.
4. As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAP; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.
5. Post ICD-10 code recommendations for public review and comment.
6. Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
7. NCQA staff finalize ICD-10 code recommendations.

*Tools Used to Identify/Map to ICD-10*

All tools used for mapping/code identification from CMS ICD-10 website

(<https://www.cms.gov/medicare/Coding/ICD10/index.html>).

GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

### *Expert Participation*

The NCQA HEDIS Expert Coding Panel and NCQA’s Behavioral Health Measurement Advisory Panel reviewed and provided feedback on staff recommendations. Names and credentials of the experts who served on these panels are listed under Additional Information, Ad. 1. Workgroup/Expert Panel Involved in Measure Development.

**2012 Submission**

NCQA identified and refined measure management into a standardized process called the HEDIS measure life cycle.

\*Step 1: Topic selection is the process of identifying measures that meet criteria consistent with the overall model for performance measurement. There is a huge universe of potential performance measures for future versions of HEDIS. The first step is identifying measures that meet formal criteria for further development.

NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure “Desirable”? The work-up is vetted by NCQA’s MAPs, the TAG, the HEDIS Policy Panel and various other panels.

\*Step 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase.

Development includes the following tasks.

1.Ensure funding throughout measure testing

2.Prepare a detailed conceptual and operational work-up that includes a testing proposal

3.Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures

The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

\*Step 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to the CPM about new measures or about changes to existing measures.

NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

\*Step 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA’s Quality Compass? Or in accreditation scoring.

The first-year distinction guarantees that a measure can be efficiently collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA’s experience is that the first year of large-scale data collection often reveals unanticipated issues.

After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

\*Step 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be reported in Quality Compass and may be used for scoring in accreditation.

Step 6: Evaluation is the ongoing review of a measure’s performance and recommendations for its modification or retirement. Every measure is reevaluated at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments contribute to measure evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, a third of the measurement set is researched for changes in clinical guidelines or health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year’s data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year’s HEDIS Volume 2.

What makes a measure “Desirable”?

Whether considering the value of a new measure or the continuing worth of an existing one, we must define what makes a measure useful. HEDIS measures encourage improvement. The defining question for all performance measurement—”Where can measurement make a difference?”—can be answered only after considering many factors. NCQA has established three areas of desirable characteristics for HEDIS measures, discussed below.

* 1. Relevance: Measures should address features that apply to purchasers or consumers, or which will stimulate internal efforts toward quality improvement. More specifically, relevance includes the following attributes.

Meaningful: What is the significance of the measure to the different groups concerned with health care? Is the measure easily interpreted? Are the results meaningful to target audiences?

Measures should be meaningful to at least one HEDIS audience (e.g., individual consumers, purchasers or health care systems). Decision makers should be able to understand a measure’s clinical and economic significance.

Important to health: What is the prevalence and overall impact of the condition in the U.S. population? What significant health care aspects will the measure address?

We should consider the type of measure (e.g., outcome or process), the prevalence of medical condition addressed by the measure and the seriousness of affected health outcomes.

Financially important: What financial implications result from actions evaluated by the measure? Does the measure relate to activities with high financial impact?

Measures should relate to activities that have high financial impact.

Cost effective: What is the cost benefit of implementing the change in the health care system? Does the measure encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness? Measures should encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness.

Strategically important: What are the policy implications? Does the measure encourage activities that use resources efficiently? Measures should encourage activities that use resources most efficiently to maximize member health.

Controllable: What impact can the organization have on the condition or disease? What impact can the organization have on the measure? Health care systems should be able to improve their performance. For outcome measures, at least one process should be controlled and have an important effect on outcome. For process measures, there should be a strong link between the process and desired outcome.

Variation across systems: Will there be variation across systems? There should be the potential for wide variation across systems.

Potential for improvement: Will organizations be able to improve performance? There should be substantial room for performance improvement.

* 1. Scientific soundness: Perhaps in no other industry is scientific soundness as important as in health care. Scientific soundness must be a core value of our health care system—a system that has extended and improved the lives of countless individuals.

Clinical evidence: Is there strong evidence to support the measure? Are there published guidelines for the condition? Do the guidelines discuss aspects of the measure? Does evidence document a link between clinical processes and outcomes addressed by the measure? There should be evidence documenting a link between clinical processes and outcomes.

Reproducible: Are results consistent? Measures should produce the same results when repeated in the same population and setting.

Valid: Does the measure make sense? Measures should make sense logically and clinically, and should correlate well with other measures of the same aspects of care.

Accurate: How well does the measure evaluate what is happening? Measures should precisely evaluate what is actually happening.

Risk adjustment: Is it appropriate to stratify the measure by age or another variable? Measure variables should not differ appreciably beyond the health care system’s control, or variables should be known and measurable. Risk stratification or a validated model for calculating an adjusted result can be used for measures with confounding variables.

Comparability of data sources: How do different systems affect accuracy, reproducibility and validity? Accuracy, reproducibility and validity should not be affected if different systems use different data sources for a measure.

* 1. Feasibility:

The goal is not only to include feasible measures, but also to catalyze a process whereby relevant measures can be made feasible.

Precise specifications: Are there clear specifications for data sources and methods for data collection and reporting? Measures should have clear specifications for data sources and methods for data collection and reporting.

Reasonable cost: Does the measure impose a burden on health care systems? Measures should not impose an inappropriate burden on health care systems.

Confidentiality: Does data collection meet accepted standards of member confidentiality?

Data collection should not violate accepted standards of member confidentiality. Logistical feasibility

Are the required data available?

Auditability: Is the measure susceptible to exploitation or “gaming” that would be undetectable in an audit? Measures should not be susceptible to manipulation that would be undetectable in an audit.

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

**2020 Submission:**

Statistical results of construct validity testing

Table 4a. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Hospitalization for Mental Illness* Performance Scores Within Measure – **Commercial** Plans, calendar year 2018 data

|  |  |
| --- | --- |
| **Rate** | **Correlation Coefficient** |
|  | 30-day |
| 7-day | 0.90\* |

\*Significant at p < 0.001

Table 4b. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Hospitalization for Mental Illness* Performance Scores Within Measure – **Medicaid** Plans, calendar year 2018 data

|  |  |
| --- | --- |
| **Rate** | **Correlation Coefficient** |
|  | 30-day |
| 7-day | 0.93\* |

\*Significant at p < 0.001

Table 4c. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Hospitalization for Mental Illness* Performance Scores Within Measure – **Medicare** Plans, calendar year 2018 data

|  |  |
| --- | --- |
| **Rate** | **Correlation Coefficient** |
|  | 30-day |
| 7-day | 0.91\* |

\*Significant at p < 0.001

Table 5a. Results of Pearson Correlation Coefficient for Commercial, Medicaid and Medicare health plans for the *Follow-Up After Hospitalization for Mental Illness* and *Follow-Up After Emergency Department Visit for Mental Illness* Measures, Calendar Year 2018 Data

|  |  |  |
| --- | --- | --- |
| Product Line | Rate | **Correlation Coefficient** |
| Commercial | FUH 7-day and FUM 7-day | .497 |
| (N=, p value =) | (316, p < 0.001) |
| FUH 30-day and FUM 30-day | .609 |
| (N=, p value =) | (316, p < 0.001) |
| FUH 7-day and FUM 30-day | .555 |
| (N=, p value =) | (316, p < 0.001) |
| FUH 30-day and FUM 7-day | .533 |
| (N=, p value =) | (316, p < 0.001) |
| Medicaid | FUH 7-day and FUM 7-day | .476 |
| (N=, p value =) | (156, p <0.001) |
| FUH 30-day and FUM 30-day | .514 |
| (N=, p value =) | (156, p <0.001) |
| FUH 7-day and FUM 30-day | .524 |
| (N=, p value =) | (156, p <0.001) |
| FUH 30-day and FUM 7-day | .452 |
| (N=, p value =) | (156, p <0.001) |
| Medicare | FUH 7-day and FUM 30-day | .537 |
| (N=, p value =) | (243, p <0.001) |
| FUH 30-day and FUM 30-day | .630 |
| (N=, p value =) | (243, p <0.001) |
| FUH 7-day and FUM 30-day | .585 |
| (N=, p value =) | (243, p <0.001) |
| FUH 30-day and FUM 30-day | .555 |
| (N=, p value =) | (243, p <0.001) |

N = the number of plans reporting both indicators.

**2016 Update**

**ICD-10 CONVERSION**

*Summary of Stakeholder Comments Received*

NCQA posted ICD-10 codes for public review and comment in March 2011 and March 2012. Comments received helped to ensure we were mapping the codes correctly.

**2012 Submission**

**Results of face validity assessment**

Step 1: The Follow-Up After Hospitalization for Mental Illness measure was developed to address a gap in care concerning follow-up care for people with mental illness. NCQA’s Performance Measurement Department and the Behavioral Health MAP worked together to assess the most appropriate tools for monitoring follow-up for mental illness.

Step 2: The measure was written, field-tested, and presented to the CPM and incorporated into HEDIS in 1994.

Step 3: The measure was released for Public Comment prior to publication in HEDIS. We received and responded to comments on this measure.

Step 4: The Follow-Up After Hospitalization for Mental Illness measure was introduced in HEDIS 1994. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year.

Step 5: The Follow-Up After Hospitalization for Mental Illness measure was reevaluated in 2011/2012.

**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?*)

**2020 Submission:**

Interpretation of construct validity testing    
Correlations between individual rates within the *Follow-Up After Hospitalization for Mental Illness* measure were positive and strong (Tables 4a, 4b, 4c) across product lines. Across all product lines, correlations between the *Follow-Up After Hospitalization for Mental Illness* and the *Follow-Up After Emergency Department Visit for Mental Illness* measure rates (Table 5a) were positive and moderate. Plans with higher rates on *Follow-Up After Hospitalization for Mental Illness* tend to also have higher rates on the *Follow-Up After Emergency Department Visit for Mental Illness* measure. The results indicate that the *Follow-Up After Hospitalization for Mental Illness* measure has good validity.

Interpretation of systematic assessment of face validity

 The multi-stakeholder advisory panels concluded the measures had good face validity.

**2016 Submission:**

**Interpretation of systematic assessment of face validity:** Our advisory panels agreed that the measures as specified will accurately differentiate quality across health plans. The measure had sufficient face validity.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#section2b4)

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

**2020 Submission:**

NCQA currently allows health plans to apply exclusions to their results. NCQA does not collect data on exclusion for HEDIS reporting of the measure. In measure development and field testing, we investigated and validated the exclusion applied to the eligible denominator.

**2016 Update: EXCLUSIONS ANALYSIS**

NCQA currently allows health plans for exclusion to their results. NCQA does not collect data on exclusion for HEDIS reporting of the measure. In measure development and field testing, we investigate and validate the exclusion applied to the eligible denominator.

**2012 Submission**

NCQA currently allows health plans for optional exclusion to their results. NCQA does not conduct the annual analysis applied to a sample. In measure development, field testing and any re-analysis for update, we investigate and validate the affect of the reliability exclusion applied to the eligible denominator.

**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*)

**2020 Submission:**

N/A

**2016 Submission:**

N/A

**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*)

**2020 Submission:**

N/A

**2016 Submission:**

N/A

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

**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*** [***2b4***](#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)*

**2020 Submission:**

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the below 25th and above 75th percentile groups. The t-test method calculates a testing statistic based on the sample size, performance rate, and standard error of each plan. The test statistic is then compared against a t distribution, which is similar to a normal distribution. If the p-value of the test statistic is less than .05, then the two plans’ performance is significantly different from each other.

**2016 Update**  
To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than .05, then the two plans’ performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance using 2016 data. We used these two plans as examples of measured entities. However the method can be used for comparison of any two measured entities.

**2012 Submission**

Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance.

Comparison of means and percentiles; analysis of variance against established benchmarks: if sample size is >400, we would use an analysis of variance.

**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*)

**2020 Submission:**

**HEDIS MY 2018 Variation in Performance across Health Plans- Commercial**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **N** | **Min** | **P10** | **P25** | **Mean** | **Median** | **P75** | **P90** | **Max** | **IQR** | **P value** |
| 7-day rate | 361 | 0 | 0.3 | 0.37 | 0.44 | 0.44 | 0.51 | 0.59 | 0.79 | 0.14 | p < 0.001 |
| 30-day rate | 358 | 0 | 0.52 | 0.6 | 0.66 | 0.67 | 0.73 | 0.78 | 0.92 | 0.13 | p < 0.001 |

**HEDIS MY 2018 Variation in Performance across Health Plans- Medicare**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **N** | **Min** | **P10** | **P25** | **Mean** | **Median** | **P75** | **P90** | **Max** | **IQR** | **P value** |
| 7-day rate | 308 | 0 | 0.13 | 0.18 | 0.28 | 0.25 | 0.34 | 0.46 | 0.68 | 0.16 | p < 0.001 |
| 30-day rate | 308 | 0.07 | 0.3 | 0.37 | 0.48 | 0.47 | 0.6 | 0.7 | 0.84 | 0.23 | p < 0.001 |

**HEDIS MY 2018 Variation in Performance across Health Plans- Medicaid**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **N** | **Min** | **P10** | **P25** | **Mean** | **Median** | **P75** | **P90** | **Max** | **IQR** | **P value** |
| 7-day rate | 173 | 0.05 | 0.21 | 0.29 | 0.36 | 0.35 | 0.43 | 0.52 | 0.7 | 0.14 | p < 0.001 |
| 30-day rate | 172 | 0.12 | 0.38 | 0.5 | 0.57 | 0.58 | 0.66 | 0.72 | 0.83 | 0.16 | p < 0.001 |

N = Number of plans reporting

IQR = Interquartile range

p-value = p-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

**2016 Update**

**HEDIS 2016 Variation in Performance across Health Plans- Commercial**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rate | # of Plans | Avg EP | Avg. | SD | Min. | 10th | 25th | 50th | 75th | 90th | IQR | P-Value |
| 7 days | 368 | 568 | 50.3% | 13.1% | 2.6% | 34.7% | 42.2% | 49.8% | 58.7% | 65.8% | 16.5% | <0.001 |
| 30 days | 368 | 568 | 69.7% | 11.1% | 7.7% | 55.4% | 64.6% | 70.6% | 76.8% | 82.5% | 12.2% | <0.001 |

EP: Eligible Population, the average denominator size across plans submitting to HEDIS

IQR: Interquartile range

p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

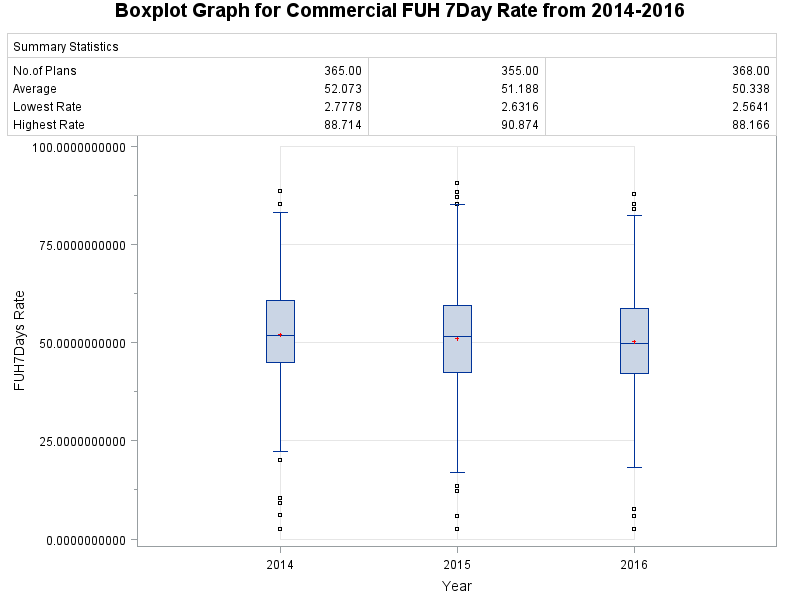
**HEDIS 2016 Variation in Performance Across Health Plans- Medicare**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rate | # of Plans | Avg. EP | Avg. | SD | Min. | 10th | 25th | 50th | 75th | 90th | IQR | P-Value |
| 7 days | 301 | 279 | 33.8% | 14.9% | 3.3% | 15.7% | 22.4% | 32.0% | 43.0% | 55.1% | 20.6% | <0.001 |
| 30 days | 301 | 279 | 52.4% | 17.0% | 11.1% | 30.6% | 39.8% | 53.5% | 65.2% | 76.2% | 25.4% | <0.001 |

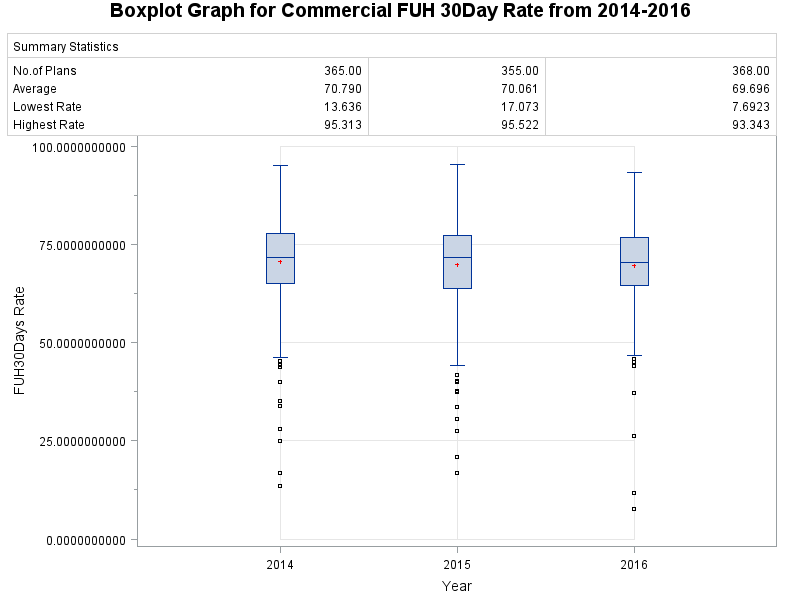
**HEDIS 2016 Variation in Performan Across Health Plans- Medicaid**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rate | # of Plans | Avg. EP | Avg. | SD | Min. | 10th | 25th | 50th | 75th | 90th | IQR | P-Value |
| 7 days | 166 | 1,182 | 43.6% | 15.7% | 0.0% | 24.7% | 34.2% | 43.6% | 55.2% | 64.2% | 21.0% | <0.001 |
| 30 days | 168 | 1,169 | 61.2% | 16.0% | 8.1% | 41.3% | 54.1% | 63.7% | 72.6% | 78.5% | 18.5% | <0.001 |

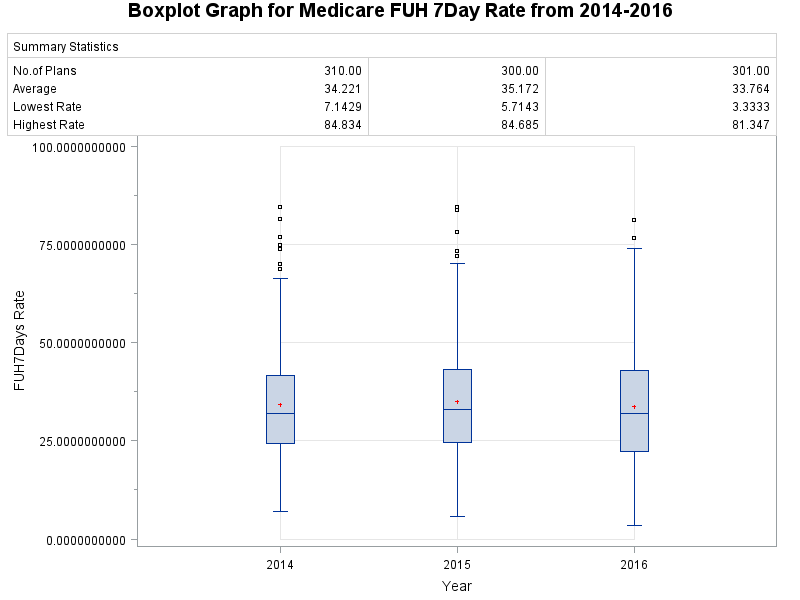
**Figure 1a. *Follow-Up After Hospitalization for Mental Illness -7-Day Rate*: Commercial Plans 2014-2016**



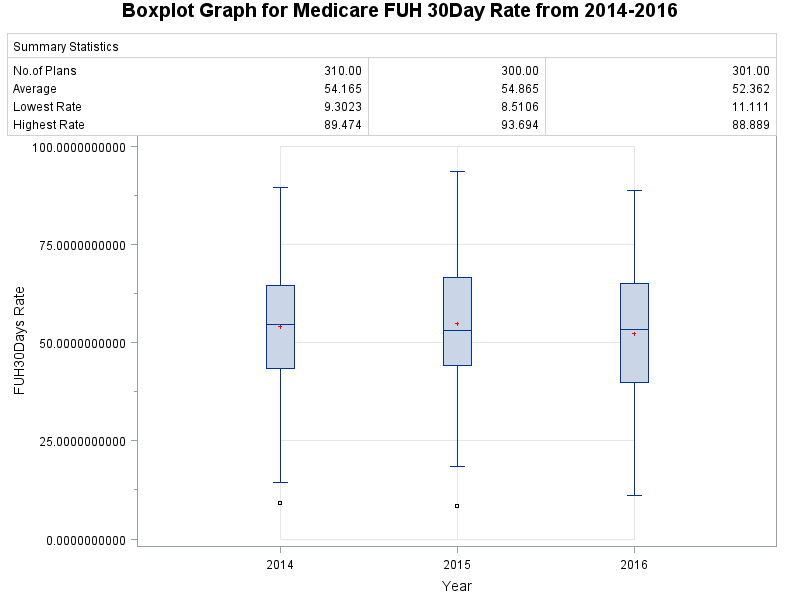
**Figure 1b. *Follow-Up After Hospitalization for Mental Illness -30-Day Rate*: Commercial Plans 2014-2016**



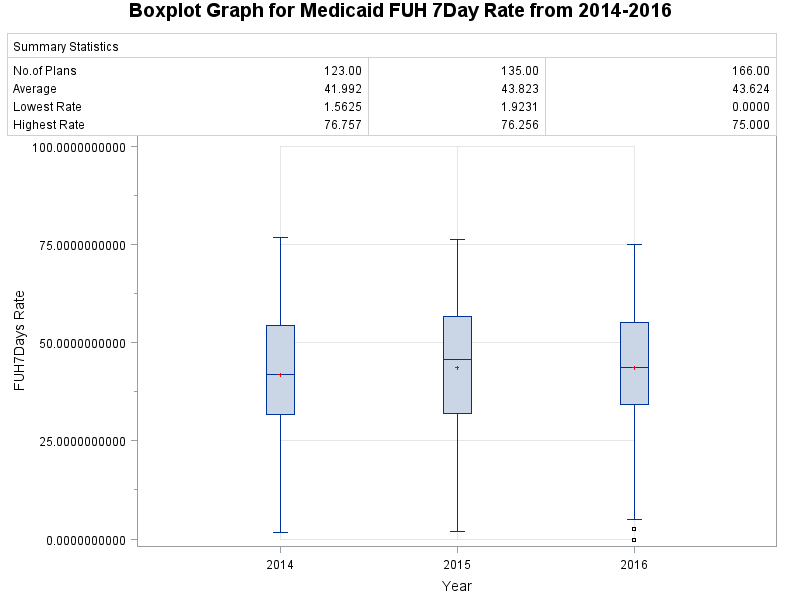
**Figure 3a. *Follow-Up After Hospitalization for Mental Illness -7-Day Rate*: Medicare Plans 2014-2016**



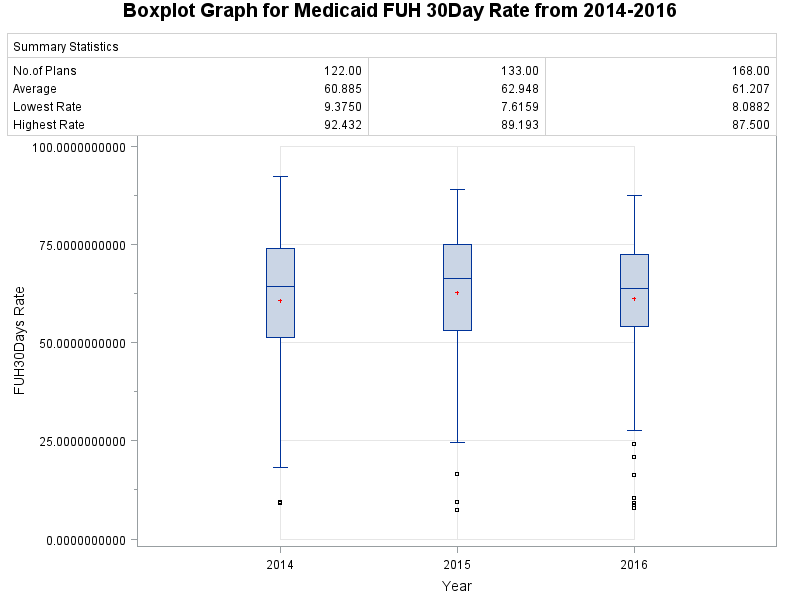
**Figure 3b. *Follow-Up After Hospitalization for Mental Illness -30-Day Rate*: Medicare Plans 2014-2016**



**Figure 2a. *Follow-Up After Hospitalization for Mental Illness -7-Day Rate*: Medicaid Plans 2014-2016**



**Figure 2b. *Follow-Up After Hospitalization for Mental Illness -30-Day Rate*: Medicaid Plans 2014-2016**



**2012 Submission**

7-Day Rate

Commercial

Measurement Year: 2009; 2010; 2011

N: 397 391 363

Min: 5.8 3.75 3.13

Max: 97.62 90.18 93.33

Mean: 54.01 55.96 57.22

SD: 13.1 13.75 12.88

P10: 37.93 39.22 42.05

P25: 45.26 46.54 48.74

P50: 53.85 56.01 57.04

P75: 62.96 65.19 66.13

P90: 71.23 72.76 72.07

Medicaid

Measurement Year: 2009; 2010; 2011

N: 62 71 85

Min: 2.6 8.2 10.87

Max: 78.57 87.9 86.85

Mean: 42.62 42.89 44.56

SD: 18.29 18.6 16.45

P10: 15.52 18.22 23.02

P25: 31.65 29.59 33.1

P50: 44.53 43.52 45.11

P75: 56.63 59.1 53.91

P90: 64.15 64.25 68.31

Medicare

Measurement Year: 2009; 2010; 2011

N: 193 231 257

Min: 4.23 2.13 1.67

Max: 86.67 84.21 84

Mean:37.97 38 37.8

SD: 17.55 18.33 18.02

P10: 15.57 13.7 15.38

P25: 23.26 23.86 24.24

P50: 36.88 36.84 37.44

P75: 51.39 50 48.45

P90: 60.32 63.49 63.93

30-Day Rate

Commercial

Measurement Year: 2009; 2010; 2011

N: 397 391 364

Min: 21.74 21.21 13.58

Max: 98.61 97.32 100

Mean: 74.1 74.68 75.93

SD: 10.31 10.8 10.49

P10: 60 61.57 64.89

P25: 67.94 68.82 71.02

P50: 74.74 76 76.38

P75: 81.82 82.21 82.43

P90: 85.96 86.29 87.2

Medicaid

Measurement Year: 2009; 2010; 2011

N: 61 70 82

Min: 18.07 15.63 22.7

Max: 87.5 91.67 87.79

Mean: 61.67 60.22 63.83

SD: 18.25 19.14 16.19

P10: 37.27 31.79 36

P25: 49.6 49.02 57.14

P50: 64.29 62.63 66.6

P75: 75.65 74.28 74.62

P90: 81.23 83.57 82.56

Medicare

Measurement Year: 2009; 2010; 2011

N: 193 230 254

Min: 9.86 5.77 5.95

Max: 100 96.15 93.33

Mean: 56.32 55.99 56.69

SD: 18.38 19.17 18.73

P10: 30 27.3 29.79

P25: 43.82 42.11 44.87

P50: 58.1 58.23 57.95

P75: 71.43 71.88 70

P90: 78.18 79.72 80

**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?*)

**2020 Submission**

For the 7-day rate, there is a 14 percentage point gap in performance between Commercial plans at the 25th and 75th percentiles, a 16 percentage point gap for Medicare plans, and a 14 percentage point gap for Medicaid plans. For the 30-day rate, there is a 13 percentage point gap in performance between Commercial plans at the 25th and 75th percentiles, a 23 percentage point gap for Medicare plans, and a 16 percentage point gap for Medicare plans. The difference in performance between plans in the 25th percentile and 75th percentile is statistically significant for both rates across all product lines.

**2016 Submission:**

The results above indicate there is a 12-25% gap in performance between the 25th and 75th performing plans. For all product lines and rates the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for the Medicare health plans which show a 20.6-25.4 percentage point gap between 25th and 75th percentile plans.

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

**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*)  
 **2020 Submission**

This measure has only one set of specifications.

**2016 Submission:**

N/A

**2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)  
**2020 Submission**

This measure has only one set of specifications.

**2016 Submission:**

N/A

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

**2020 Submission**

This measure has only one set of specifications.

**2016 Submission:**

N/A

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

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

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

**2020 Submission**

HEDIS measures apply to enrolled members in a health plan, and NCQA has a rigorous audit process to ensure the eligible population and numerator events for each measure are correctly identified and reported. The audit process is designed to verify primary data sources used to populate measures and ensure specifications are correctly implemented.

The HEDIS Compliance Audit addresses the following functions:

* Information practices and control procedures
* Sampling methods and procedures
* Data integrity
* Compliance with HEDIS specifications
* Analytic file production
* Reporting and documentation

**2016 Submission:**

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to missing data.

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

**2020 Submission: 2020 Submission**

HEDIS addresses missing data in a structured way through its audit process. HEDIS measures apply to enrolled members in a health plan, and NCQA-certified auditors use standard audit methodologies to assess whether data sources are missing data. If a data source is found to be missing data, and the issues cannot be rectified, the auditor will assign a “materially biased” designation to the measure for that reporting plan, and the rate will not be used. Once measures are added to HEDIS, NCQA conducts a first-year analysis to assess the feasibility of the measure when widely implemented in the field. This analysis includes an assessment of how many plans report valid rates vs. rates that are materially biased (or have other issues, such as small denominators). These considerations are weighed in the deliberation process before measures are approved for public reporting.

**2016 Submission:**

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to missing data.

**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*)

**2020 Submission**

This measure goes through the NCQA audit process each year to identify potential errors or bias in results. Only performances rates that have been reviewed and determined not to be “materially biased” are reported and used.

**2016 Submission:**

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to missing data.