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

**Measure Number** (*if previously endorsed*)**:** NQF #2456

**Measure Title**: Medication Reconciliation: Number of Unintentional Medication Discrepancies per Patient

**Date of Submission**: 8/1/2019

**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.   *\*No risk adjustment or risk 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: Medication data collected from patient/caregiver interview, ambulatory providers, community pharmacies, electronic prescription fill information and information on discrepancies in medication orders (intentional or not) from provider interviews. | other: Medication data collected from patient/caregiver interview, ambulatory providers, community pharmacies, electronic prescription fill information and information on discrepancies in medication orders (intentional or not) from provider interviews. |

**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*).

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

December 2011 – June 2014

**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*)

Five U.S. hospitals were selected, based on their willingness to participate in the MARQUIS study and ability to collect data and implement interventions to improve the medication reconciliation process. These included two academic medical centers, two community hospitals (one teaching, one non-teaching), and one Veterans Affairs medical center. These hospitals were geographically diverse (West coast, Northeast, Southeast, and Midwest) and varied in size from 45 to 653 beds. Two utilized an electronic health record prior to data collection, two implemented an EHR during the data collection period, and one continued to use a paper medical record supplemented by medication reconciliation software.

**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*)

For testing the reliability of gold-standard medication histories, we included 19 randomly selected medical inpatients at one large, urban, academic medical center.

For testing the reliability of the discrepancy scoring system, we included 4 patients, one each from 4 of the MARQUIS study sites, chosen by study pharmacists because of the challenging nature of the cases.

**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**.

**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.

Not applicable. There were no social risk factors analyzed.

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

**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*)

To evaluate inter-rater reliability of the gold-standard medication histories, 19 randomly selected

medication histories were collected independently by two study pharmacists. Among all the medications recorded for each patient, there was complete agreement in medication, dose, route, and frequency for 147 of 192 medications (77%).

To evaluate inter-rater reliability of the discrepancy scoring system, we analyzed the last 4 quarterly cases, consisting of a total of 44 medications and 128 ratings each for admission and discharge discrepancies (i.e., 256 data points). We found the following:

* For the presence of admission discrepancies, we found agreement for 116/128 ratings (91% agreement)
* For the presence of discharge discrepancies, we found agreement for 116/128 ratings (91% agreement)
* When an admission discrepancy was present (according to the gold standard), we found complete agreement for discrepancy type (i.e., omission, dose, frequency, route, formulation, additional, other) in 55/64 cases (86%), partial agreement (e.g., “dose and frequency” vs. “dose”) in 4/64 cases (6%), and disagreement in 5/64 cases (8%)
* When an admission discrepancy was present, we found complete agreement for discrepancy reason (i.e., history error vs. reconciliation error) in 47/64 cases (73%).
* When a discharge discrepancy was present, we found complete agreement for discrepancy type in 46/56 cases (82%), partial agreement in 4/56 cases (7%) and disagreement in 6/56 cases (11%).
* When a discharge discrepancy was present, we found complete agreement for discrepancy reason in 45/56 cases (80%).

The kappa for the presence of admission discrepancies was 0.64 (substantial agreement) and the kappa for the presence of discharge discrepancies was also 0.64 (substantial agreement) across all raters.

**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*)  
In the analysis of reliability of the scoring system, kappas were statistically significant from zero (for admission discrepancies, Z=7.29, p<0.0001; for discharge discrepancies, Z=7.34, p<0.0001).

**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?*)

Inter-rater reliability for the gold-standard medication history was high.

Inter-rater reliability for the presence or absence of an unintentional discrepancy (i.e., the information required to calculate a discrepancy rate per patient) was substantial. Reliability was somewhat lower for discrepancy type and reason, but these are to be used for internal QI purposes only (i.e., are not part of the measure itself). It should be noted that the cases used to derive these analyses were chosen because they were challenging for study pharmacists to evaluate – reliability for an “average” case is likely higher.

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

**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**

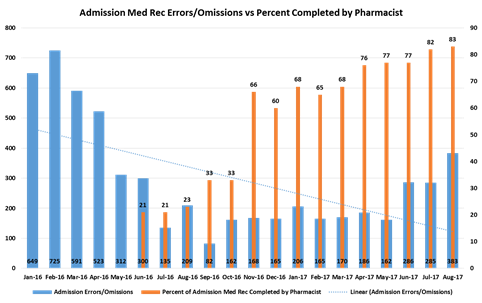
**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing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

The literature supports that pharmacists take more accurate medication histories than either nurses or physicians. (1) It is therefore reasonable to assume that a specially trained pharmacist taking a preadmission medication history is a good proxy for a gold-standard medication history. Comparing errors in admission and discharge orders compared to this gold standard has high face validity as a measure of the quality of medication reconciliation. After all the goal of medication reconciliation is “to identify the most accurate list of all medications a patient is taking . . . and using this list to provide correct medications for patients anywhere within the health care system.” In other words, what matters clinically is whether the orders are correct, based on knowing what medications the patient was taking previously. No other measure in existence looks at this process more directly than the measure proposed here.

The process for measuring discrepancies is systematic and transparent – see the Leapfrog worksheet and workbook for details on the process, now being used at hundreds of Leapfrog hospitals. Those who measure discrepancies are trained experts (trained pharmacists) at each site. See the attached materials for how they are trained and the above paragraph on how pharmacists can justifiably be considered medication experts. Performance scores, or at least relative improvement in scores over time, can be used to distinguish sites that truly improved from those that did not. For example, in the MARQUIS2 study, involving 17 sites, 9 sites had significant improvement in their discrepancy rates per medication per patient in the last 6 months of the study compared with the first 6 months. Compared with those sites that did not show improvement, those that did show improvement had a greater increase in the proportion of patients who received patient-level interventions (55% absolute improvement vs. 22% absolute improvement), such as a best-possible medication history taken by a dedicated trained provider while the patient was still in the Emergency Department (see below).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **First 6 months** | | **Last 6 months** | | **Relative Risk (95% CI)** | **P value** | **Proportion of Patients who received patient-level interventions** | | |
|  | **N** | **Discrepancies per medication per patient** | **N** | **Discrepancies per medication per patient** |  |  | **First 6 months** | **Last 6 months** | **Absolute improvement** |
| Site 1 | 54 | 0.24 | 161 | 0.12 | **0.50** | **<0.0001** | **24%** | **82%** | **58%** |
| Site 2 | 174 | 0.37 | 287 | 0.26 | **0.70** | **<0.0001** | **56%** | **68%** | **13%** |
| Site 3 | 31 | 0.41 | 222 | 0.17 | **0.41** | **<0.0001** | **17%** | **59%** | **42%** |
| Site 4 | 136 | 0.24 | 98 | 0.19 | 0.79 | 0.11 | 44% | 50% | 6% |
| Site 5 | 94 | 0.26 | 159 | 0.27 | 1.04 | 0.64 | 33% | 45% | 12% |
| Site 6 | 110 | 0.24 | 268 | 0.26 | 1.08 | 0.21 | 61% | 95% | 34% |
| Site 7 | 68 | 0.16 | 459 | 0.23 | **1.44** | **0.0004** | **2%** | **61%** | **59%** |
| Site 8 | 39 | 0.08 | 149 | 0.04 | **0.50** | **0.01** | **1%** | **91%** | **89%** |
| Site 9 | 29 | 0.46 | 319 | 0.32 | **0.70** | **0.0001** | **1%** | **80%** | **79%** |
| Site 10 | 72 | 0.40 | 124 | 0.26 | **0.65** | **<0.0001** | **0%** | **86%** | **86%** |
| Site 11 | 56 | 0.32 | 294 | 0.31 | 0.97 | 0.63 | 20% | 35% | 15% |
| Site 12 | 45 | 0.47 | 82 | 0.36 | **0.77** | **0.003** | **37%** | **53%** | **15%** |
| Site 13 | 48 | 0.30 | 288 | 0.30 | 1.00 | 0.98 | 44% | 34% | -9% |
| Site 14 | 47 | 0.08 | 100 | 0.11 | 1.38 | 0.17 | 0% | 28% | 28% |
| Site 15 | 84 | 0.34 | 223 | 0.22 | **0.65** | **<0.0001** | **0%** | **100%** | **100%** |
| Site 16 | 102 | 0.17 | 225 | 0.14 | 0.82 | 0.053 | 0% | 28% | 28% |
| Site 17 | 40 | 0.15 | 259 | 0.10 | **0.67** | **0.01** | **38%** | **48%** | **10%** |

These results can also be seen in Leapfrog data. For example, when a particular site started having pharmacists take medication histories, their number of discrepancies went down:



References:

1. Dawson P, Gray S. Clinical significance of pharmacist-obtained drug histories. Pharm J. 1981;227:120.

**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)*

Face validity: as explained above, the measure looks directly at medication orders and compares them to a gold-standard medication history, then determines whether these discrepancies are intentional or not based on the medical record +/- provider interview. No other measure in existence looks at this process more directly than the measure proposed here.

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

Not applicable.

**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?*)

Not applicable.

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

**2b2. EXCLUSIONS ANALYSIS**

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

Patients may be excluded from the measure if they are discharged or expire prior to being seen by a study pharmacist, are otherwise unavailable to be seen by a pharmacist, or decline to talk to the pharmacist.

**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*)

In 2006 at Brigham and Women’s Hospital, we compared the 180 patients included in the analysis with the 199 excluded subjects. As expected, compared with excluded subjects, study patients were older, had longer lengths of stay, and had more medications at discharge. (1) In this study, patients were required to provide informed written consent, so the differences between included and excluded patients may have been more pronounced than if this were a routine part of hospital measurement.

References:

1. Pippins JR, Gandhi TK, Hamann C, et al. Classifying and predicting errors of inpatient medication reconciliation. J Gen Intern Med. 2008;23(9):1414-1422.

(Reprint attached – see Table 1)

This is an unusual measure, in that there really is no better measure of the quality of medication reconciliation with which is conduct a comparison.  As shown in our systematic review, interventions that reduce medication discrepancies as measured using our methodology have been shown to improve more downstream patient outcomes, including potential adverse drug events, actual drug events, and (occasionally) post-discharge health care utilization.

References:

1. Mueller SK, Sponsler KC, Kripalani S, Schnipper JL. Hospital-based medication reconciliation practices: a systematic review. Arch Intern Med. 2012;172(14):1057-1069. PMID: 22733210.

To certify that pharmacists have mastery over the measurement process, “Gold standard pharmacists” receive the following in terms of training and certification:

* Review of a slide presentation and video on how to take a gold standard medication
* Completion of a pre- and post-test on the fundamentals of history-taking
* Satisfactory completion of at least one simulated case (>90% accuracy of gold standard history, >90% of best practice behaviors demonstrated)

We have created an online platform to administer these components and track completion of these requirements.  The simulated cases themselves are available in the MARQUIS “train the trainer” materials: <https://shm.hospitalmedicine.org/acton/media/25526/shm---bphm-train-the-trainers-material>.

The training materials on being a gold standard pharmacist are here: <https://shm.hospitalmedicine.org/acton/media/25526/shm-data-pharmacist-training-part-1>

In addition, by definition, we require anyone taking a gold-standard history to have pharmacist-level credentials because studies have shown that pharmacists take more accurate medication histories than physicians or nurses

Lastly, as part of Leapfrog, we review the results of each site on a regular basis and talk with sites with outlier results to make sure they understand the measure and correct any misunderstanding of how to measure it.

**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*)

These exclusions are simply practical ones (i.e., there is no alternative)

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

**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**.

Certain patients are at higher risk for medication discrepancies than others, but since this is an intermediate patient outcome (a measure of processes of care), risk stratification is less critical than for the evaluation of more distal outcomes such as adverse drug events. In the name of simplicity, we have chosen not to recommend risk adjustment.

However, because the number of discrepancies in medication orders is highly correlated with the total number of gold standard medications (each medication is an opportunity for error), we suggest modifying the previous metric (or adding a second metric) to be the number of unintentional medication discrepancies per medication per patient. The maximum number of this metric is two: there is a discrepancy in the admission order and a discrepancy in the discharge order for each medication.

References:

1. Pippins JR, Gandhi TK, Hamann C, et al. Classifying and predicting errors of inpatient medication reconciliation. J Gen Intern Med. 2008;23(9):1414-1422.
2. Salanitro AH, Osborn CY, Schnipper JL, Roumie CL, Labonville S, Johnson DC, Neal E, Cawthon C, Businger A, Dalal AK, Kripalani S. Effect of patient- and medication-related factors on inpatient medication reconciliation errors. J Gen Intern Med. Aug 2012;27(8):924-932.

**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-Leeson 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)*

We used the two-sample T test method (1-2), using PASS power calculation software (PASS 2008, NSCC LLC, Kaysville, Utah), assuming an alpha of 0.05, beta of 0.20 (i.e., 80% power), baseline results and standard errors from the MARQUIS study, and effect sizes that are both achievable and close to the smallest sizes that would be considered clinically meaningful.

References:

1. Machin, D., Campbell, M., Fairs, P., and Pinal, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, MA.

Zarf, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

**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*)

Using data from MARQUIS, we can assume a baseline rate of 3.44 medication discrepancies per patient with a standard deviation of 3.62. With an alpha of 0.05, then with 6 months of pre-intervention data collection (at 25 patients per month, or 150 patients total) and 12 months of post-intervention data collection (300 patients), hospitals would have 80% power to detect a reduction from 3.44 to 2.43 discrepancies per patient.

Using data from the MARQUIS2 study using unintentional medication discrepancies per medication per patient (5022 patients at 18 hospitals), the baseline rate (i.e., among patients who did not receive interventions) was 0.62, with a standard deviation of 0.14 per site. With an alpha of 0.05, then with 6 months of pre-intervention data collection (at 25 patients per month, or 150 patients total) and 12 months of post-intervention data collection (300 patients), hospitals would have 80% power to detect a reduction from 0.62 to 0.58 discrepancies per medication per patient, a 6.5% relative reduction, close to the smallest effect size that could be considered clinically important. This effect size was seen by all 18 of the participating MARQUIS2 sites during the study period among patients who received interventions compared with those who did not.

**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?*)

With data collection on approximately 1 patient per weekday and a reasonable time-frame, hospitals will have adequate statistical power to detect effect sizes in discrepancy rates per patient that are achievable (see above for our results from the MARQUIS2 study) and close to the smallest effect size that could be considered clinically important.

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

**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 embrasures). 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 nonresponse’s) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

To minimize sampling bias, we used a method whereby the list of admitted patients from the day prior are randomized in the order in which they are approached. Once approached, the goal is for the pharmacist to take a gold-standard medication history before the patient is discharged from the hospital unless the patient declines. This may require multiple attempts (e.g., because the patient is off the floor at a procedure or test or wants the pharmacist to return at another time). By approaching patients on the first full day after admission, this method also minimizes bias by length of stay. As noted above (**2b2.2.**), patients who were measured were generally older, with longer lengths of stay, and on more medications, but these differences would likely be less in a non-research setting. We know of no better way to minimize selection bias for this metric.

Once selected for measurement, there should be no missing data, as all data collection is inherent to the process of taking a gold-standard medication history and comparing it to medication orders, which should always be accessible, and reviewing the medical record to determine whether discrepancies were unintentional, which should also always be available.

**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*)

As noted above, for the metric of number of unintentional medication discrepancies (per patient or per medication per patient), there is essentially no missing data. Reasons for discrepancies (e.g., history errors vs. reconciliation errors) was missing in X% of cases for the MARQUIS2 study.

**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*)

We believe our method minimizes selection bias to the greatest extent that is practical. As noted above, missing data is not an issue with this measure.

**Addendum**

Key Developments Since the Approval of the Measure

Since our metric was originally approved by NQF, there have been two major developments:

1. We completed the second Multicenter Medication Reconciliation Quality Improvement Study (MARQUIS2), which used this metric to measure the quality of medication reconciliation and response to interventions among over 5000 patients in 18 diverse hospitals, making it the largest medication reconciliation interventional study conducted to date in the U.S.
2. The Leapfrog Group adopted this metric, with the only modification, based on our advice, of dividing the total number of unintentional medication discrepancies by the total number of gold standard medications (i.e., opportunities for error). This metric has now been reported by 1,123 hospitals, including 21,347 patients, over this past year (see table below). The number of discrepancies per medication per patient was fairly stable over the last two years, with a fairly wide spread in performance among reporting hospitals. In this process of measure adoption, The Leapfrog Group has also created instructional materials and data collection tools to ease the burden of collection and reporting, held public forums, addressed questions from sites, and taken other measures to maximize consistent adoption of this metric.

**Table: Preliminary Results of Leapfrog Group**

|  |  |  |  |
| --- | --- | --- | --- |
| **Measurement Year** | **Number of Hospitals** | **Number of Patients** | **Discrepancies per medication per patient, median (IQR)** |
| 2017 | 980 | 12,291 | 0.16 (0.07-0.31) |
| 2018 | 1,123 | 21,347 | 0.15 (0.06-0.28) |