



## Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

### Brief Measure Information

**NQF #:** 3746

**Corresponding Measures:**

**Measure Title:** Avoid Hospitalization After Release with a Misdiagnosis—ED Stroke/Dizziness (Avoid H.A.R.M.—ED Stroke/Dizziness)

**Measure Steward:** Johns Hopkins Armstrong Institute for Patient Safety and Quality

**sp.02. Brief Description of Measure:**

This outcome measure tracks the rate of adult patients (aged 18 years and older) treated and released from the Emergency Department (ED) with either a non-specific, presumed benign symptom-only dizziness diagnosis or a specific inner ear/vestibular diagnosis (collectively referred to as "benign dizziness") who were subsequently admitted to a hospital for a stroke within 30 days of their ED visit.

The measure accounts for the epidemiologic base rate of stroke in the population under study using a risk difference approach (observed [short-term rate, reflecting days 0-30 days] minus expected [long-term rate, reflecting days 91-360]).

**1b.01. Developer Rationale:**

---

**sp.12. Numerator Statement:** The number of ED treat-and-release index visit discharges during the performance period that are followed within 30 days by an inpatient hospital admission to any hospital that ends in a primary discharge diagnosis of stroke.

**sp.14. Denominator Statement:** Patients treated and released from the ED with a primary discharge diagnosis code of "benign dizziness". A patient's first such discharge during the performance period will be considered the "index visit". Any subsequent ED treat-and-release discharge with a diagnosis of 'benign dizziness' that falls outside a 360-day follow-up window from the previous qualifying "index visit" will be considered another distinct "index visit".

**sp.16. Denominator Exclusions:** The measure has no exclusions. All patients treated and released from the ED with "benign dizziness" as their primary discharge diagnosis code are included in the measure denominator

---

**Measure Type:** Outcome: Intermediate Clinical Outcome

**sp.28. Data Source:**

Claims

**sp.07. Level of Analysis:**

Facility

---

**IF Endorsement Maintenance – Original Endorsement Date:**

**Most Recent Endorsement Date:**

---

**IF this measure is included in a composite, NQF Composite#/title:**

**IF this measure is paired/grouped, NQF#/title:**

**sp.03. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?:**

## 1. Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria

---

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Measure and Report: Evidence section. For example:

**Current Submission:**

Updated evidence information here.

**Previous (Year) Submission:**

Evidence from the previous submission here.

**1a.01. Provide a logic model.**

*Briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.*

**[Response Begins]**

**[Response Ends]**

**1a.02. Select the type of source for the systematic review of the body of evidence that supports the performance measure.**

*A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data.*

**[Response Begins]**

**[Response Ends]**

If the evidence is not based on a systematic review, skip to the end of the section and do not complete the repeatable question group below. If you wish to include more than one systematic review, add additional tables by clicking "Add" after the final question in the group.

**Evidence - Systematic Reviews Table (Repeatable)**

Group 1 - Evidence - Systematic Reviews Table

**1a.03. Provide the title, author, date, citation (including page number) and URL for the systematic review.**

**[Response Begins]**

**[Response Ends]**

**1a.04. Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the systematic review.**

[Response Begins]

[Response Ends]

**1a.05. Provide the grade assigned to the evidence associated with the recommendation, and include the definition of the grade.**

[Response Begins]

[Response Ends]

**1a.06. Provide all other grades and definitions from the evidence grading system.**

[Response Begins]

[Response Ends]

**1a.07. Provide the grade assigned to the recommendation, with definition of the grade.**

[Response Begins]

[Response Ends]

**1a.08. Provide all other grades and definitions from the recommendation grading system.**

[Response Begins]

[Response Ends]

**1a.09. Detail the quantity (how many studies) and quality (the type of studies) of the evidence.**

[Response Begins]

[Response Ends]

**1a.10. Provide the estimates of benefit, and consistency across studies.**

[Response Begins]

[Response Ends]

**1a.11. Indicate what, if any, harms were identified in the study.**

[Response Begins]

[Response Ends]

**1a.12. Identify any new studies conducted since the systematic review, and indicate whether the new studies change the conclusions from the systematic review.**

[Response Begins]

[Response Ends]

**1a.13. If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, describe the evidence on which you are basing the performance measure.**

[Response Begins]

[Response Ends]

**1a.14. Briefly synthesize the evidence that supports the measure.**

[Response Begins]

[Response Ends]

**1a.15. Detail the process used to identify the evidence.**

[Response Begins]

[Response Ends]

**1a.16. Provide the citation(s) for the evidence.**

[Response Begins]

[Response Ends]

**1b.01. Briefly explain the rationale for this measure.**

*Explain how the measure will improve the quality of care, and list the benefits or improvements in quality envisioned by use of this measure.*

[Response Begins]

[Response Ends]

**1b.02. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.**

*Include mean, std dev, min, max, interquartile range, and scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.*

[Response Begins]

[Response Ends]

**1b.03. If no or limited performance data on the measure as specified is reported above, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement. Include citations.**

[Response Begins]

[Response Ends]

**1b.04. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.**

*Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Include mean, std dev, min, max, interquartile range, and scores by decile. For measures that show high levels of performance, i.e., “topped out”, disparities data may demonstrate an*

*opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.*

**[Response Begins]**

**[Response Ends]**

**1b.05. If no or limited data on disparities from the measure as specified is reported above, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in above.**

**[Response Begins]**

**[Response Ends]**

## 2. Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.

### sp.01. Provide the measure title.

Measure titles should be concise yet convey who and what is being measured (see [What Good Looks Like](#)).

#### [Response Begins]

Avoid Hospitalization After Release with a Misdiagnosis—ED Stroke/Dizziness (Avoid H.A.R.M.—ED Stroke/Dizziness)

#### [Response Ends]

### sp.02. Provide a brief description of the measure.

Including type of score, measure focus, target population, timeframe, (e.g., Percentage of adult patients aged 18-75 years receiving one or more HbA1c tests per year).

#### [Response Begins]

This outcome measure tracks the rate of adult patients (aged 18 years and older) treated and released from the Emergency Department (ED) with either a non-specific, presumed benign symptom-only dizziness diagnosis or a specific inner ear/vestibular diagnosis (collectively referred to as “benign dizziness”) who were subsequently admitted to a hospital for a stroke within 30 days of their ED visit.

The measure accounts for the epidemiologic base rate of stroke in the population under study using a risk difference approach (observed [short-term rate, reflecting days 0-30 days] minus expected [long-term rate, reflecting days 91-360]).

#### [Response Ends]

### sp.04. Check all the clinical condition/topic areas that apply to your measure, below.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- Surgery: General

#### [Response Begins]

Neurology: Stroke/Transient Ischemic Attack (TIA)

#### [Response Ends]

### sp.05. Check all the non-condition specific measure domain areas that apply to your measure, below.

#### [Response Begins]

Other (specify)

[Other (specify) Please Explain]

Safety: Diagnostic error

[Response Ends]

**sp.06. Select one or more target population categories.**

*Select only those target populations which can be stratified in the reporting of the measure's result.*

*Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.*

*Please do not select:*

- *Populations at Risk: Populations at Risk*

[Response Begins]

Adults (Age >= 18)

[Response Ends]

**sp.07. Select the levels of analysis that apply to your measure.**

*Check ONLY the levels of analysis for which the measure is SPECIFIED and TESTED.*

*Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.*

*Please do not select:*

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]

Facility

[Response Ends]

**sp.08. Indicate the care settings that apply to your measure.**

*Check ONLY the settings for which the measure is SPECIFIED and TESTED.*

[Response Begins]

Ambulatory Care

[Response Ends]

**sp.09. Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials.**

*Do not enter a URL linking to a home page or to general information. If no URL is available, indicate "none available".*

[Response Begins]

[https://www.hopkinsmedicine.org/armstrong\\_institute/centers/center\\_for\\_diagnostic\\_excellence/dizzy-stroke-ed-specs.html](https://www.hopkinsmedicine.org/armstrong_institute/centers/center_for_diagnostic_excellence/dizzy-stroke-ed-specs.html)

**[Response Ends]**

**sp.12. Attach the data dictionary, code table, or value sets (and risk model codes and coefficients when applicable). Excel formats (.xlsx or .csv) are preferred.**

*Attach an excel or csv file; if this poses an issue, [contact staff](#). Provide descriptors for any codes. Use one file with multiple worksheets, if needed.*

**[Response Begins]**

Available in attached Excel or csv file

**[Response Ends]**

Attachment: 3746\_Avoid HARM Dizzy-Stroke ICD-10 Codes.xlsx

**sp.13. State the numerator.**

*Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome).*

*DO NOT include the rationale for the measure.*

**[Response Begins]**

The number of ED treat-and-release index visit discharges during the performance period that are followed within 30 days by an inpatient hospital admission to any hospital that ends in a primary discharge diagnosis of stroke.

**[Response Ends]**

**sp.14. Provide details needed to calculate the numerator.**

*All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets.*

*Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.*

**[Response Begins]**

For each patient's index ED visit identified in the denominator, identify if the patient had an inpatient hospital admission to any hospital within 30 days of their ED discharge date that resulted in a primary diagnosis of stroke. The ICD-10 codes to be used to identify patients with a primary diagnosis of stroke can be found in the submitted Excel file.

**[Response Ends]**

**sp.15. State the denominator.**

*Brief, narrative description of the target population being measured.*

**[Response Begins]**

Patients treated and released from the ED with a primary discharge diagnosis code of “benign dizziness”. A patient’s first such discharge during the performance period will be considered the “index visit”. Any subsequent ED treat-and-release discharge with a diagnosis of ‘benign dizziness’ that falls outside a 360-day follow-up window from the previous qualifying “index visit” will be considered another distinct ‘index visit’.

**[Response Ends]**

**sp.16. Provide details needed to calculate the denominator.**

*All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets.*

*Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.*

**[Response Begins]**

Using a 36- month performance period, identify those ED patients who were treated and released from the ED with a primary discharge diagnosis of “benign dizziness”. This includes patients with either with (1) a specific benign dizziness diagnosis (e.g., benign paroxysmal positional vertigo) or (2) a non-specific, symptom-only dizziness diagnosis (i.e., dizziness or vertigo, not otherwise specified). The ICD-10 codes to be used to identify patients with a primary diagnosis of “benign dizziness” can be found in the submitted Excel file.

A patient’s first ED treat-and-release discharge during the performance period meeting the above criteria should be included in the denominator. This is considered the patient’s first “index visit”. A patient’s second “index visit” is the first subsequent ED treat-and-release discharge meeting the above criteria that is more than 360 days after the first index visit’s ED discharge date and this “index visit” should also be included in the denominator. A patient’s third “index visit” is the first subsequent ED treat-and-release discharge meeting the above criteria that is more than 360 days after the second index visit’s ED discharge date and this “index visit” should be included in the denominator. A patient’s fourth “index visit” is the first subsequent ED treat-and-release discharge meeting the above criteria that is more than 360 days after the third index visit’s ED discharge date and this “index visit” should be included in the denominator.

The denominator value is the count of the number of ED “index visits” with a primary discharge diagnosis of “benign dizziness” during the performance period. The maximum number of “index visits” for a single patient in a 36-month performance period is 4.

**[Response Ends]**

**sp.17. Describe the denominator exclusions.**

*Brief narrative description of exclusions from the target population.*

**[Response Begins]**

The measure has no exclusions. All patients treated and released from the ED with "benign dizziness" as their primary discharge diagnosis code are included in the measure denominator

**[Response Ends]**

**sp.18. Provide details needed to calculate the denominator exclusions.**

*All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.*

**[Response Begins]**

Not applicable.

**[Response Ends]**

**sp.19. Provide all information required to stratify the measure results, if necessary.**

*Include the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate. Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format in the Data Dictionary field.*

**[Response Begins]**

Not applicable.

**[Response Ends]**

**sp.20. Is this measure adjusted for socioeconomic status (SES)?**

**[Response Begins]**

No

**[Response Ends]**

**sp.21. Select the risk adjustment type.**

*Select type. Provide specifications for risk stratification and/or risk models in the Scientific Acceptability section.*

**[Response Begins]**

No risk adjustment or risk stratification

**[Response Ends]**

**sp.22. Select the most relevant type of score.**

*Attachment: If available, please provide a sample report.*

**[Response Begins]**

Rate/proportion

**[Response Ends]**

**sp.23. Select the appropriate interpretation of the measure score.**

*Classifies interpretation of score according to whether better quality or resource use is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score*

**[Response Begins]**

Better quality = Lower score

**[Response Ends]**

**sp.24. Diagram or describe the calculation of the measure score as an ordered sequence of steps.**

*Identify the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period of data, aggregating data; risk adjustment; etc.*

**[Response Begins]**

Steps to calculate an ED's risk of misdiagnosed-related harm from missed stroke.

a. Step 1 – Identify all patients treated and released from the ED with a primary discharge diagnosis of “benign dizziness” during the 36-month performance period.

b. Step 2– A patient's first ED discharge during the 36-month performance period with a primary discharge diagnosis of “benign dizziness” should be included in the denominator. This patient discharge is considered the patient's first “index visit”. A patient's (potential) second “index visit” is the first subsequent ED treat-and-release visit with a discharge diagnosis of “benign dizziness” that is more than 360 days after the first index visit's ED discharge date. A patient's (potential) third “index visit” is the first subsequent ED treat-and-release visit with a discharge diagnosis of “benign dizziness” that is more than 360 days after the second index visit's ED discharge date. A patient's (potential) fourth “index visit” is the first subsequent ED treat-and-release visit with a discharge diagnosis of “benign dizziness” that is more than 360 days after the third index visit's ED discharge date. Index visits that do not have patients enrolled for at least 360 days after the index visit should be excluded.

c. Step 3 – Count the number of ED “index visits”—this is the denominator value. The maximum number of “index visits” for a single patient in a 36-month performance period is 4.

**“Observed” Rate Calculation**

d. Step 4 – For each “index visit” in Step 3, identify if the patient had an inpatient admission to any hospital within 30 days of their ED index visit discharge that resulted in a primary discharge diagnosis of stroke. Count the number of “index visits” that meet this criterion—this is the short-term 30-day numerator value for incident strokes.

e. Step 5 – Measure the observed rate. Crude short-term 30-day incidence rate per 10,000 visits = (Step 4: [number of short-term stroke hospitalizations within 30d + alpha] / Step 3: [number of eligible ED benign dizziness discharges in the performance period + 1]) x 10,000. The constants “alpha” = 1/1,000 (for the numerator) and “1” (for the denominator) are added to avoid issues with possible zero counts [see footnote “\*” below for clarification].

**“Expected” Rate Calculation**

f. Step 6 – For each “index visit” in Step 3, identify if the patient had an inpatient admission to any hospital with a primary discharge diagnosis of stroke in the time window 91 days through 360 days following their ED index visit discharge. Count the number of “index visits” that meet this criterion. This is the measured numerator value for long-term incident strokes.

g. Step 7 – Divide the number of strokes identified in Step 6 (from 91 days through 360 days) by 9 to obtain a ‘monthly’ value. This is the long-term 30-day-equivalent (i.e., monthly average) numerator value for incident strokes. This is needed to calculate the average long-term stroke incidence rate per 30 days.

h. Step 8 – Measure the expected rate. Crude long-term 30-day incidence rate per 10,000 visits = (Step 7: [average number of long-term stroke hospitalizations per 30d + alpha] / Step 4: [number of eligible ED benign dizziness treat-and-release discharges in the performance period who did not experience a stroke in the prior 90 days + 1 - (3 x alpha)]) x 10,000. The denominator should exclude those patients who experienced a stroke prior to 90 days as we are only counting the first stroke in the 91-360 days post index visit. The constants “alpha” = 1/1,000 (for the numerator) and “1 - (3 x alpha)” (for the denominator) are added to avoid issues with possible zero counts [see footnote “\*” below for clarification].

**“Attributable” Rate (Measure) Calculation**

i. Step 9 – Attributable 30d rate per 10,000 visits = Step 5 (crude short-term 30d rate) – Step 8 (crude long-term 30d rate)

\* The constants “alpha” = 1/1,000 (for the numerator) and “1” (for the denominator) are added to avoid issues with possible zero counts. This is equivalent to a posterior estimation using Beta (alpha, 1-alpha) as prior for each 30-day rate. It is similar to the “add 0.5” approach in the Fisher’s exact test with low counts, except that here, the 30-day stroke return rate of alpha = 1/1,000 is used as prior as opposed to 1/2 as in the Fisher’s exact test. This prior translates to adding 1 observation with a 30-day stroke return rate of alpha when calculating the observed 30-day rate and the expected 30-day rate. The estimation is asymptotically unbiased and consistent. The effect of this statistical adjustment is negligible but penalizes the measure towards no harm. The statistical adjustment factor (alpha) of 1/1,000 was chosen to be similar to the long-term, baseline stroke risk after ED treat-and-release discharge (~0.1%) and is reasonable based on our current data and that from prior research studies. Removing “3 x alpha” from the denominator in calculating the expected 30d rate is due to having to remove patients that already experienced a stroke hospitalization prior to 90d.

**[Response Ends]**

**sp.27. If measure testing is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.**

*Examples of samples used for testing:*

- Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.
- The sample should represent the variety of entities whose performance will be measured. The [2010 Measure Testing Task Force](#) recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.
- The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.
- When possible, units of measurement and patients within units should be randomly selected.

**[Response Begins]**

Not applicable

**[Response Ends]**

**sp.30. Select only the data sources for which the measure is specified.**

**[Response Begins]**

Claims

**[Response Ends]**

**sp.31. Identify the specific data source or data collection instrument.**

*For example, provide the name of the database, clinical registry, collection instrument, etc., and describe how data are collected.*

**[Response Begins]**

Not applicable

**[Response Ends]**

**sp.32. Provide the data collection instrument.**

**[Response Begins]**

No data collection instrument provided

**[Response Ends]**

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate fields in the Scientific Acceptability sections of the Measure Submission Form.

- 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.
- All required sections must be completed.
- For composites with outcome and resource use measures, Questions 2b.23-2b.37 (Risk Adjustment) also must be completed.
- If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), Questions 2b.11-2b.13 also must be completed.
- An appendix for supplemental materials may be submitted (see Question 1 in the Additional section), but there is no guarantee it will be reviewed.
- Contact NQF staff with any questions. Check for resources at the [Submitting Standards webpage](#).
- 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 the [2021 Measure Evaluation Criteria and Guidance](#).

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.

2a. Reliability testing 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 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;

AND

If patient preference (e.g., informed decision-making) 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).

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; 14,15 and has demonstrated adequate discrimination and calibration
- 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 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 non-responders) and how the specified handling of missing data minimizes bias.

2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:

2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and

2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.

(if not conducted or results not adequate, justification must be submitted and accepted)

## Definitions

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).

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 measure 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.

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.

Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

Risk factors that influence outcomes should not be specified as exclusions.

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.

Please separate added or updated information from the most recent measure evaluation within each question response in the Scientific Acceptability sections. For example:

**Current Submission:**

Updated testing information here.

**Previous (Year) Submission:**

Testing from the previous submission here.

**2a.01. Select only the data sources for which the measure is tested.**

**[Response Begins]**

Claims

**[Response Ends]**

**2a.02. 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).*

**[Response Begins]**

**Calculating and testing the performance measure:** For this analysis, we used two different data sources with complementary strengths and relative weaknesses to highlight the potential reliability and validity of the performance measure under different circumstances. The ***national hospital-level analysis (using Medicare data)*** provides a comprehensive assessment of all hospitals in the US (strength) but represents predominantly older adults >65yo (relative weakness). It also ensures virtually complete capture of hospital crossovers (i.e., ED index visit treat-and-release discharge at hospital A followed later by an inpatient hospitalization for stroke at hospital B), even if these crossovers occur across health systems or across state lines (strength) but misses a large fraction of relevant ED index visits (i.e., patients aged 18-64yo), lowering measure precision (relative weakness). The ***state hospital-level analysis (using HCUP data)*** offers a more limited range of hospitals that may not be fully representative of all hospitals in the US (relative weakness) but represents adults of all ages >18yo (strength). Although it may miss some out-of-state hospital admission crossovers (relative weakness), it captures all relevant ED index visits at each of the included hospitals (i.e., adults of any age), improving measure precision (strength).

**National hospital-level testing:** This analysis used de-identified national Medicare Fee-for-Service (FFS) Parts A & B claims and enrollment data (*approved for reuse under CMS DUA RSCH-2020-55692*) in combination with de-identified administrative claims data and enrollment data from the OptumLabs® Data Warehouse (OLDW), selecting members of Medicare Advantage (MA) plans.

**State hospital-level testing:** This analysis used de-identified state-level Inpatient administrative claims (SID) and Emergency Department administrative claims (SEDD) for Florida hospitals, as made available through the Agency for Healthcare Research and Quality's Healthcare Utilization Project (HCUP).

**Data element validity testing:** This analysis used a combination of electronic health record (EHR) data and associated claims data from the four Johns Hopkins Health System hospitals in Maryland (two academic medical centers and two community hospitals).

**[Response Ends]**

**2a.03. Provide the dates of the data used in testing.**

Use the following format: “MM-DD-YYYY - MM-DD-YYYY”

**[Response Begins]**

01-01-2015 -12-31-2019

**[Response Ends]**

**2a.04. Select the levels of analysis for which the measure is tested.**

Testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- Clinician: Clinician
- Population: Population

**[Response Begins]**

Facility

**[Response Ends]**

**2a.05. List the measured entities 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.

**[Response Begins]**

**National hospital-level testing:**

Using the Medicare FFS data (but not Medicare Advantage data from OLDW), we identified facilities that had at least one claim with a CPT code of 9928x during the performance period, indicating that facility billed for an ED visit. This filter identified 5,503 unique facilities which appears to be a reasonable capture of all hospital-based EDs in the United States since there are just over 6,100 hospitals in the U.S. (*AHA Fast Facts 2020, based on FY2018 AHA Survey data*) and some hospitals do not systematically care for Medicare patients (e.g., Department of Defense hospitals).

OptumLabs used the facility IDs identified through our “CPT 9928x filter” and identified the number of ED visits at each facility as recorded in the 2017 AHA Survey data. The aggregate distribution of ED visits at the identified hospitals matched well with the 2010 study by Muelleman et al. (*Acad Emerg Med*) that looked at ED visit volume distribution across U.S. hospitals (with expected growth in visits during the last 10 years).

ED Visits per year	Muelleman et al. (2007 data) N=4,874 Non-Federal EDs	Our Dataset (2017 data) N=5,503 Medicare EDs
<10,000	31%	32%
10,000-19,999	21%	16%
20,000-29,000	15%	12%
30,000-39,999	13%	10%

ED Visits per year	Muelleman et al. (2007 data) N=4,874 Non-Federal EDs	Our Dataset (2017 data) N=5,503 Medicare EDs
40,000-49,000	8%	8%
>50,000	12%	23%

**Table 1. Comparison of ED Visits between 2010 study and our National Hospital Dataset.**

For the measure analysis, we used 967 of the 5,503 facilities. These 967 facilities had at least 250 “benign dizziness” treat-and-release ED discharges during the 3-year performance period and therefore were likely to have a large enough sample size to produce a reliable measure of performance. Hospitals with 250 “benign dizziness” treat-and-release discharges in Medicare data typically reflect medium to larger hospital EDs that see roughly 40,000 to 50,000 ED visits per year (depending on patient demographic mix and insurance mix).

Due to data privacy constraints, we could not access descriptive statistics on the 967 facilities used in the measure analysis. These 967 facilities (17.6% of the total 5,503) are presumably disproportionately those EDs with higher numbers of annual visits. Besides the obvious characteristics of larger EDs (e.g., located in larger population centers), there could be differences related to access to technology or specialists that decrease the likelihood of error. We do not anticipate any additional systematic biases involving the facilities included in the analysis.

**State hospital-level testing:**

The HCUP SEDD data for Florida identified 216 unique EDs that were included in our state-level testing. This number reflects 98% of the 220 non-federal, short-term, acute care hospitals in Florida ([American Hospital Directory - Individual Hospital Statistics for Florida \(ahd.com\)](#))

The aggregate distribution of visits in Florida EDs skewed a bit higher than the 2010 study by Muelleman et al. (*Acad Emerg Med*) that looked at ED visit volume distribution across U.S. hospitals, but this could be a function of national growth in ED visits during the last 10 years and/or the general population growth in Florida since that time.

ED Visits per year	Muelleman et al. (2007 data) N=4,874 Non-Federal EDs	Our Dataset (2016-2019 data) N=2016 Florida EDs
<10,000	31%	8%
10,000-19,999	21%	21%
20,000-29,000	15%	22%
30,000-39,999	13%	18%
40,000-49,000	8%	13%Tab
>50,000	12%	18%

**Table 2. Comparison of ED Visits between 2010 study and our State Hospital Dataset.**

For the measure testing, we used all 216 facilities.

Due to data privacy constraints, we could not access descriptive statistics on the 216 facilities used in the measure analysis. These 216 facilities, however, are likely representative of all 220 hospital-based EDs in the state of Florida.

**[Response Ends]**

**2a.06. Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis), separated by level of analysis and data source; if a sample was used, describe how patients were selected for inclusion in the sample.**

*If there is a minimum case count used for testing, that minimum must be reflected in the specifications.*

**[Response Begins]**

**National hospital-level testing:**

A total of 1,232,389 ED treat-and-release visits with a “benign dizziness” discharge diagnosis were included in the testing and analysis. These reflect treat-and-release discharges from the 967 hospital EDs during the 3-year performance period. The age distribution is as expected for Medicare data. The female-to-male distribution is typical for dizziness across age groups (roughly 60% female, 40% male).

Patient Demographics of ED Treat-and-Release Visits with a “Benign Dizziness” Discharge Diagnosis	Percentage of Patients (%)
Age	
• 18-24	0.19%
• 25-44	3.49%
• 45-59	8.11%
• 60-74	40.15%
• 75+	48.06%
• Unknown	0.00%
Sex	
• Male	38.36%
• Female	61.64%
• Unknown	0.00%
Race/Ethnicity	
• White	74.66%
• Black/African-American	12.80%
• Asian/Pacific Islander	2.88%
• Hispanic	7.59%
• Other/Unknown	2.07%

**Table 3. Percentage of Patients in National Hospital Dataset with Demographic Characteristic.**

**State hospital-level testing:**

A total of 208,472 ED treat-and-release visits with a “benign dizziness” discharge diagnosis were included in the testing and analysis. These reflect treat-and-release discharges from the 216 hospital EDs during the 3-year performance period. Note that the age distribution is skewed slightly older than national populations with dizziness in the ED (PMID: 18613993), as expected for Florida, which has a higher percentage of residents over age 65 than any state other than Maine (<https://www.prb.org/resources/which-us-states-are-the-oldest/>). The female-to-male distribution is typical for dizziness across age groups (roughly 60% female, 40% male).

Patient Demographics of ED Treat-and-Release Visits with a “Benign Dizziness” Discharge Diagnosis	Percentage of Patients (%)
Age	
• 18-24	5.77%
• 25-44	25.18%
• 45-59	24.94%
• 60-74	24.99%

Patient Demographics of ED Treat-and-Release Visits with a “Benign Dizziness” Discharge Diagnosis	Percentage of Patients (%)
• 75+	16.94%
• Unknown	0.00%
Sex	
• Male	37.20%
• Female	62.80%
• Unknown	0.00%
Race/Ethnicity	
• White	54.17%
• Black/African-American	20.69%
• Asian/Pacific Islander	1.13%
• Hispanic	21.31%
• Other/Unknown	0.11%

**Table 4. Percentage of Patients in State Hospital Dataset with Demographic Characteristic.**

**[Response Ends]**

**2a.07. 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.**

**[Response Begins]**

**National score-level reliability testing (Medicare FFS and Medicare Advantage from OLDW):** Data from January 1, 2015 – December 31, 2017 were used for the score-level reliability testing and variation in performance across hospitals.

**State score-level reliability testing (Florida HCUP data):** Data from January 1, 2016 – December 31, 2019 were used for the score-level reliability testing and variation in performance across hospitals.

**Data-element validity testing (Johns Hopkins Health System):** Data from July 1, 2016 – June 30, 2017 were used for data-element validity testing.

**[Response Ends]**

**2a.08. List 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.*

**[Response Begins]**

No social risk factors were available or directly analyzed. However, our risk difference approach (“observed minus expected”) that accounts for baseline stroke risk accounts for social determinants of long-term stroke risk in the cohort of patients who are at risk and being measured. Although some social risk factors likely impact the risk of misdiagnosis (e.g., patients who identified their race as Black or African-American are more likely to have their stroke misdiagnosed (PMID: 28344918), it would be inappropriate to “adjust” this away—if an institution

systematically performs worse in diagnosing Black or African-American patients and cares for more of these patients than the average hospital, this should not be “evened out” to match an “average” population with an “average” proportion of Black/African-American patients.

**[Response Ends]**

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.09 check patient or encounter-level data; in 2a.010 enter “see validity testing section of data elements”; and enter “N/A” for 2a.11 and 2a.12.

**2a.09. Select the level of reliability testing conducted.**

*Choose one or both levels.*

**[Response Begins]**

Accountable Entity Level (e.g., signal-to-noise analysis)

**[Response Ends]**

**2a.10. For each level of reliability testing 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.*

**[Response Begins]**

Performance measure score reliability was calculated using signal-to-noise analysis as described in the technical report by J.L. Adams titled “The Reliability of Provider Profiling: A Tutorial” (RAND Corporation, TR-653-NCQA, 2009), where the signal is the proportion of variability in measured performance that can be explained by real differences in performance. In this context, reliability represents the ability of a measure to confidently distinguish the performance of one facility from another.

**[Response Ends]**

**2a.11. For each level of reliability testing checked above, what were the statistical results from reliability testing?**

*For example, provide the percent agreement and kappa for the critical data elements, or distribution of reliability statistics from a signal-to-noise analysis. For score-level reliability testing, when using a signal-to-noise analysis, more than just one overall statistic should be reported (i.e., to demonstrate variation in reliability across providers). If a particular method yields only one statistic, this should be explained. In addition, reporting of results stratified by sample size is preferred (pg. 18, [NQF Measure Evaluation Criteria](#)).*

**[Response Begins]**

***National hospital-level testing***

We plotted a histogram of the reliability scores for the 967 facilities included in the national sample.

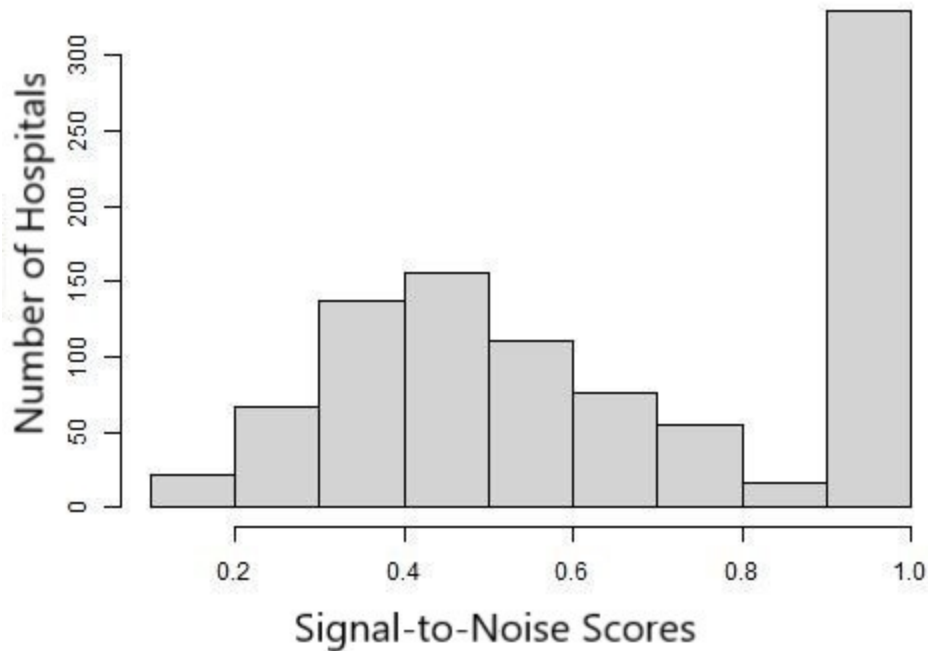


Figure 1. Histogram of Signal-to-Noise Reliability Scores for National Hospital-Level Testing

The median reliability score for the entire 967-hospital sample was 0.590, with an interquartile range of 0.414-0.951.

We also stratified our sample by the number of “benign dizziness” treat-and-release discharges in the 3-year performance window to look at the median reliability score for each stratum. As expected, reliability was higher when the number of visits analyzed was higher.

Number of Medicare “Benign Dizziness” Treat-and-Release Discharges in the 3-Year Performance Window	Median Reliability Score
250-499	0.582
500-749	0.710
750+	0.807

Table 5. Median Reliability Scores Stratified by Number of Medicare “Benign Dizziness” Treat-and-Release Visits.

**State hospital-level testing**

We plotted a histogram of the reliability scores for the 216 facilities included in the state sample.

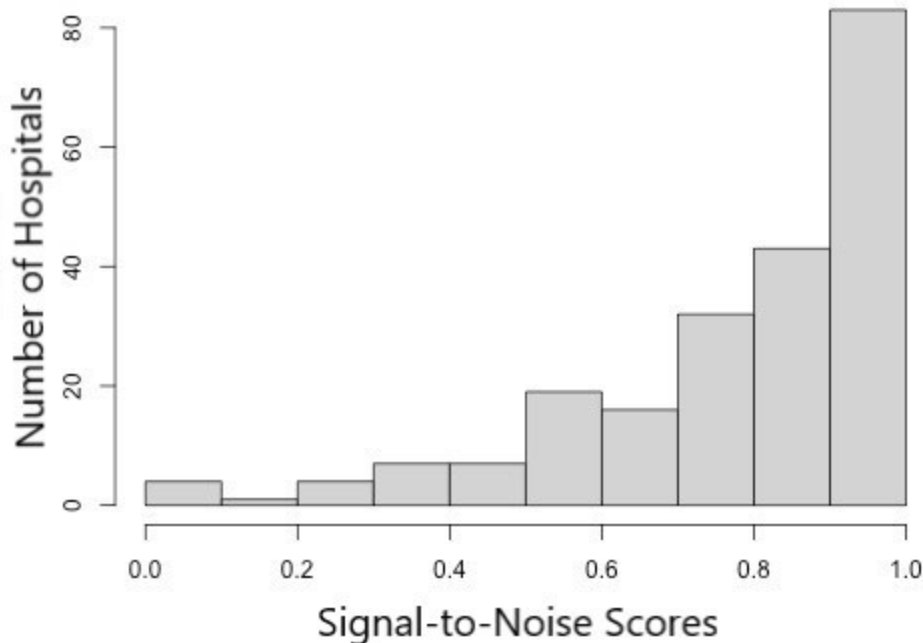


Figure 2. Histogram of Signal-to-Noise Reliability Scores for State Hospital-Level Testing

The median reliability score for the entire 216-hospital sample was 0.853, with an interquartile range of 0.671-0.950. As expected, reliability was much higher in the state-level analysis than in the national-level analysis, because of data missingness in Medicare data (i.e., it represents only ~25% of eligible ED index visits, largely because of the age constraint).

**[Response Ends]**

**2a.12. Interpret the results, in terms of how they demonstrate reliability.**

*(In other words, what do the results mean and what are the norms for the test conducted?)*

**[Response Begins]**

Reliability scores vary from 0.0 to 1.0, with a score of zero indicating that all variation is attributable to measurement error (noise, or variation across patients within the accountable entity) whereas a reliability of 1.0 implies that all variation is caused by real difference in performance across accountable entities. The reliability score depends on the pool of facilities that are included in the sample, and the reliability score is unique to each facility in that pool.

While there is not a clear cut-off for a minimum reliability level, a median value very close to 0.60 is considered by many to be sufficient for seeing differences between some entities. For the national hospital-level testing we did, which included only Medicare ED treat-and-release visits (representing only ~25% of the measure-eligible ED index visits at each facility), the smallest facilities included in the analysis (those with 250-499 “benign dizziness” treat-and-release discharges in the 3 year performance period) saw a median reliability score value of 0.582, which is very close to the 0.60 threshold previously mentioned. When we did state hospital-level testing of the measure, using HCUP data (which includes 100% of measure-eligible ED treat-and-release discharges at each facility), the

median reliability score improved to 0.853, which is well above the 0.6 threshold. Even the lower bound of the interquartile range had a reliability score of 0.67, indicating good reliability for more than three quarters of all hospitals in Florida. In other words, when data are available on all measure-eligible ED index visits, as is the case when using the Florida HCUP data, the reliability of the measure is excellent.

**[Response Ends]**

**2b.01. Select the level of validity testing that was conducted.**

**[Response Begins]**

Patient or Encounter-Level (data element validity must address ALL critical data elements)

**[Response Ends]**

**2b.02. 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.*

**[Response Begins]**

***Measure numerator (patients with an inpatient hospitalization with a diagnosis of stroke)***

Three key studies have previously evaluated the validity of using administrative data to identify stroke discharges from acute care hospitals in the U.S by comparing discharge codes against chart abstraction as the gold standard.

1. Tirschwell et. al. (*Stroke*, 2002; PMID: 12364739) looked at stroke hospitalizations for patients aged 20-years or older in Seattle, Washington, hospitals, identified by using the Comprehensive Hospital Abstract Reporting System, years 1990-1996 (N=147). Inpatient ICD-9-CM codes included 430 for intracranial hemorrhage and 431 for subarachnoid hemorrhage. Codes for ischemic stroke included 433.x1, 434, (excluding 434.x0) and 436. Cases were excluded if they had a traumatic brain injury (ICD-9-CM 800-804, 850-854), or were admitted for rehabilitation care (primary ICD-9-CM code V57). The claims-based ICD codes evaluated by Tirschwell et al. in their study have a strong overlap with the ICD codes that this measure's specifications are based on.

2. McCormick et al. (*PLoS One*, 2015; PMID: 26292280) conducted a systematic review of studies reporting on the validity of International Classification of Diseases (ICD) codes for identifying stroke in administrative data. They searched MEDLINE and EMBASE for studies prior to February 2015 that met these criteria: (a) used administrative data to identify stroke or (b) evaluated the validity of stroke codes in administrative data; and (c) reported validation statistics (sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), or Kappa scores) for stroke, or data sufficient for their calculation. Additional articles were located by hand search. Studies solely evaluating codes for transient ischemic attack were excluded. Data were extracted by two independent reviewers; article quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies tool. Positive predictive value is a measure of criterion validity. Also known as a measure of precision, it is defined here as the proportion of records with a given ICD-9-CM code that when compared with chart abstraction (the gold standard) are found to have the correct coded diagnosis for stroke. Sensitivity is a measure of the proportion of coded records which are correctly identified as such. Specificity is a measure of the proportion of records that are not coded as stroke which are correctly identified as not having a stroke. Sensitivity and specificity are closely related to the concepts of type I and type II errors.

3. A study by Kokotailo and Hill (*Stroke*, 2005; PMID: 16020772) compared hospital discharge abstract coding using ICD-9 and ICD-10 for stroke in three Canadian hospitals (one academic medical center, two community hospitals). The study authors independently reviewed a random 717 stroke patients charts that were coded using ICD-9 (charts from April 2000 to March 2001) and 249 stroke patient charts that were coded using ICD-10 (charts from April 2002 to March 2003). Using a before-and-after time period design, they compared the accuracy of hospital coding of stroke using ICD-9 and ICD-10.

**Measure denominator (patients treated and released from the ED with a discharge diagnosis of "benign dizziness")**

**Part A**

For dizziness (denominator = ED "benign dizziness" treat-and-release visit discharges), we conducted two studies focused on code-level reliability/validity.

**Question #1 (Positive Predictive Value): If an ED patient is coded with a "benign dizziness" discharge diagnosis code, how often do charts suggest the ED provider INTENDED to code a "benign dizziness" discharge diagnosis?**

Data Sources: Data from four Johns Hopkins Health System hospitals (JHHS) were used for this analysis, including two academic medical centers and two community hospitals. Data were pulled from the EPIC EHR (i.e., ICD diagnosis codes [derived from both hospital facility fee & professional fee coded diagnoses], chief complaints, and ED chart notes).

Performance Period: Jul 1, 2016 – Jun 30, 2017

Analysis: We began with a census of all cohort cases for this portion of the analysis. We stratified this group into three subgroups, based on the nature of their ED Index Visit Epic chief complaint:

- Dizziness chief complaint (dizziness/vertigo)
- Oto-vestibular chief complaint (ataxia/gait disturbance, nausea/vomiting, hearing loss/tinnitus, or ear pain)
- Other chief complaint

The dizziness chief complaint subgroup was assumed to have a valid (true positive) benign dizziness discharge diagnosis, as their presenting symptoms matched their discharge diagnosis. We did not review these charts manually. For the other two groups, we manually reviewed charts to determine whether the "benign dizziness" code was unintended (i.e., miscoded). Each chart was reviewed independently by one emergency physician and one neuro-otologist; disagreements were resolved through discussion or adjudication by a third reviewer, if necessary. This consensus opinion was judged to represent the original ED provider's intent and was used as the reference standard for determining validity.

We calculated the PPV of the ICD-10-CM codes for the entire cohort and subgroups:

$PPV = (\text{true positives}) / \text{all positives}$

Calculations are based on data from all four JHHS hospitals collectively with a stratified sampling scheme based on hospitals to ensure each hospital contributed adequate samples. We reviewed a random sub-sample of 64 charts for each non-dizziness sub-group to estimate the positive predictive value (PPV) of the benign dizziness discharge codes.

**Part B**

**Question #2 (Negative Predictive Value): If an ED patient is coded with something OTHER than a "benign dizziness" discharge diagnosis code, how often do charts suggest the ED provider INTENDED to code something OTHER than a "benign dizziness" discharge diagnosis?**

Data Sources: Data from four Johns Hopkins Health System hospitals (JHHS) were used for this analysis, including two academic medical centers and two community hospitals. Data were pulled from the EPIC EHR (i.e., ICD diagnosis codes; chief complaints; ED chart notes)

Performance Period: Jul 1, 2016 – Jun 30, 2017

Analysis Plan: We began with a census of all cohort cases for this portion of the analysis. We stratified this group into two subgroups based on the nature of their ED Index Visit Epic chief complaint and additional discharge diagnoses:

- High-risk for misclassification of "not dizziness" (Boolean 'OR' for all three criteria listed below --- i.e., "a OR b OR c")
  - a) ED (Epic) structured chief complaint of dizziness/vertigo at ED Index Visit triage
  - b) Benign dizziness diagnosis (HCUP CCS 6.8.2) in a non-primary position at ED Index Visit

c) Middle (as opposed to inner) ear diagnosis (HCUP CCS 6.8.3) in any position at ED Index Visit

- Low-risk for misclassification of “not dizziness” (all others)

The low-risk for misclassification subgroup was assumed to have a valid (true negative) not benign dizziness discharge diagnosis since their presenting symptoms matched their discharge diagnosis. We did not review these charts manually. We manually reviewed charted records for the high-risk for misclassification group to determine whether the “not benign dizziness” code was unintended (i.e., miscoded). Each chart was reviewed independently by one emergency physician and one neuro-otologist; disagreements were resolved through discussion or adjudication by a third reviewer, if necessary. This consensus opinion was judged to represent the original ED provider’s intent and was used as the reference standard for determining validity.

We calculated the NPV of the ICD-10-CM codes for the entire cohort and subgroups:

$NPV = (\text{true negatives}) / \text{all negatives}$

Calculations are based on data from all four JHHS hospitals collectively with a stratified sampling scheme based on hospitals to ensure that each hospital contributed adequate samples. We reviewed a random sub-sample of 67 charts for the high-risk sub-group to estimate the negative predictive value (NPV) of the “not benign dizziness” discharge codes.

### **Discharge Status**

Only ED patients with a disposition status of “Discharged” are included in the measure’s denominator. To confirm that ED patients with a “Discharged” disposition status were actually discharged from the ED to home, we reviewed 25 random ED patient charts from the four Johns Hopkins Health System hospitals that had a “Discharged” status between July 2016 and June 2017. We did not review any patient charts with a status other than “Discharged” as experience tells us that opportunity for misclassification of ED patients with a disposition status of “Left Against Medical Advice” or “Screened & Left” is very low since those patients typically need to complete paperwork releasing the hospital of liability before they leave the facility. We further reviewed a high-risk subset of cases from the numerator (discharged to “observation” or “clinical decision unit” rather than full hospital admission, and those with a next-day stroke admission) to make sure that they were, indeed, discharged from the ED in the first place at the ED index visit.

**[Response Ends]**

### **2b.03. Provide the statistical results from validity testing.**

*Examples may include correlations or t-test results.*

**[Response Begins]**

#### **Measure numerator (patients with an inpatient hospitalization with a diagnosis of stroke)**

In general, ICD-coded diagnoses for stroke are extremely accurate at the level of granularity required for this measure (i.e., any true cerebrovascular event case, regardless of subtype). Their accuracy drops off as higher levels of granularity are demanded (e.g., whether the stroke is an ischemic or hemorrhagic stroke). In addition, most stroke codes reflect very high specificity with fairly high (but lower) sensitivity. Key results from the articles mentioned above are as follows:

1. In the Tirschwell study (PMID: 12364739), the sensitivity for ischemic stroke was 86% (95% CI; 73–94), specificity was 95% (95% CI; 88–98), and the positive predictive value was 90% (95% CI; 77–97) with a kappa agreement score of 0.82. For intracranial hemorrhage, the sensitivity was 82% (95% CI 66–92), specificity was 93% (95% CI 86–97), and the positive predictive value was 80% (95% CI 64–91) with a kappa score of 0.74. For subarachnoid hemorrhage, the sensitivity was 98% (95% CI 90–100), specificity was 92% (95% CI 84–96), and the positive predictive value was 86% (95% CI 75–94) with a kappa score of 0.87.

- The McCormick systematic review (PMID: 26292280) included 77 published manuscripts between 1976–2015. The sensitivity of ICD-9 430-438/ICD-10 I60-I69 for any cerebrovascular disease was  $\geq 82\%$  in most [ $\geq 50\%$ ] studies, and specificity and NPV were both  $\geq 95\%$ . The PPV of these codes for any cerebrovascular disease was  $\geq 81\%$  in most studies while the PPV specifically for acute ischemic stroke, subarachnoid, or intracerebral hemorrhages (as opposed to transient ischemic attacks, other brain hemorrhages, or other cerebrovascular diseases) was  $\leq 68\%$ . In at least 50% of studies, PPVs were  $\geq 93\%$  for subarachnoid hemorrhage (ICD-9 430/ICD-10 I60), 89% for intracerebral hemorrhage (ICD-9 431/ICD-10 I61) and 82% for ischemic stroke (ICD-9 434/ICD-10 I63 or ICD-9 434&436).
- In the Kokotailo and Hill study (PMID: 16020772) they found that stroke coding was equally good with ICD-9 (90% correct [95% CI 86-93]) and ICD-10 [92% correct (95% CI 88-95)]. There were some differences in coding by stroke type, notably with transient ischemic attack, but these differences were not statistically significant.

**Measure denominator (patients treated and released from the ED with a discharge diagnosis of "benign dizziness")**

**Part A**

If the true PPV is 98% or above, a sample size of 32 gives 85% power to reject the null hypothesis that the PPV is 85% or below. The estimated PPVs and their 95% confidence intervals are summarized in the table below.

Performance Period and Chief Complaint (CC) Categories	Number of ED Index Visits	Number of Matched Records	Proportion Estimates of Matched Records	95% Confidence Intervals
JHHS – Jul 2016 – Jun 2017	1826			
CC dizziness	1308	1308/1308*	100%*	99.72-100%*
CC oto-vestibular	97	32/32	100%	89.11-100%
CC other	421	32/32	100%	89.11-100%
<b>TOTAL</b>	1826	1372/1372	100%	99.89-100%

Table 6. Positive Predictive Values for "Benign Dizziness" Discharge Diagnosis.

\* These charts were not manually reviewed but were matched based on an Epic-recorded dizziness chief complaint.

**Part B**

If the true NPV is 95% or above, a sample size of 67 gives 85% power to reject the null hypothesis that the NPV is 85% or below. The estimated NPVs and their 95% confidence intervals are summarized in the table below.

Performance Period and Risk Categories	Number of ED Index Visits	Number of Matched Records	Proportion Estimates of Matched Records	95% Confidence Intervals
JHHS – Jul 2016 – Jun 2017	99464			
High risk group	12744	66/67	98.51%	91.96-99.96%
Low risk group	86720	86720/86720*	100%*	99.996-100%*
<b>TOTAL</b>	99464	86786/86787	99.997%	99.993-99.999%

Table 7. Negative Predictive Values for "Benign Dizziness" Discharge Diagnosis.

\* These charts were not manually reviewed but were matched based on absence of any dizziness chief complaint, benign dizziness diagnosis in any position, or middle ear diagnosis in any position in the electronic Epic record.

**Discharge Status**

100% of the 25 ED charts that were reviewed with a "Discharged" disposition status were found to have an accurate status. 100% of the 6 high-risk ED patient charts in the numerator were found to have accurate status (3

were discharged from the ED to observation/clinical decision units and returning with stroke hospitalizations within days 1-30 after post-observation ED discharge; 3 were same-day return hospitalizations after treat-and-release ED discharge). Despite the fact that 3 of 3 same-day return hospitalizations were true discharges, our measure conservatively excludes potential same-day hospital admissions to avoid confusion about discharge status, even if the dataset indicates a discharge followed by an admission.

[Response Ends]

**2b.04. Provide 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?)**

[Response Begins]

***Measure numerator (patients with an inpatient hospitalization with a diagnosis of stroke)***

Both the Tirschwell and the McCormick studies found the sensitivity, specificity, and positive predictive values of the ICD-9 stroke codes to be very high (85%+) and higher still when considering accuracy as a “cerebrovascular event.” It is important to note that for most of their analyses, they demanded a higher degree of granularity in stroke diagnosis than our measure requires (e.g., if a brain hemorrhage was coded as an ischemic stroke in their study, it would have been counted as miscoded and counted against coding accuracy measures, despite being correctly coded as a “stroke” hospitalization event for our measure). These findings give us confidence about using claims data to identify patients who have had a primary stroke diagnosis for their inpatient admission. The Kokotailo and Hill study found that ICD-9 and ICD-10 were similarly accurate in capturing stroke diagnoses in three Canadian hospitals, giving us confidence that the ICD-10 coding system is useful for capturing numerator events.

***Measure denominator (patients treated and released from the ED with a discharge diagnosis of “benign dizziness”)***

We found a positive predictive value (PPV) of 100% [CI: 99.89%-100.00%] for coding “benign dizziness.” Of the 64 charts reviewed (and 1,308 electronically confirmed), all of the ED treat-and-release visit patients coded with a “benign dizziness” discharge diagnosis had a charted record that suggested that the ED provider intended to code “benign dizziness” as the discharge diagnosis. This included oversampling of high-risk charts for manual review. This gives us confidence that the codes we have outlined for identifying “benign dizziness” patients are indeed capturing encounters in which the provider intended for that diagnosis.

We found a negative predictive value (NPV) of 99.997% [CI: 99.993-99.999%] for coding “not benign dizziness.” Of the 67 charts reviewed (and 86,720 electronically confirmed) all but 1 that were coded as “not benign dizziness” had a charted record that suggested that the ED provider intended to code “not benign dizziness” as the discharge diagnosis. This included oversampling of high-risk charts for manual review. Given the high NPV (99.9%+), we feel confident that the coding is valid to support an accurate denominator (i.e., that we are not missing many cases of “true” benign dizziness among all discharges).

***Discharge Status***

The audit we completed of the “Discharged” disposition status of ED patients at the four hospitals indicates that the “Discharged” status appears to be a valid indicator of the patient’s actual discharge disposition (100% accuracy, CI: 88.8-100%), even in the highest-risk cases.

[Response Ends]

**2b.05. 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 in Importance to Measure and Report: Gap in Care/Disparities.*

**[Response Begins]**

We undertook two strategies to understand if there are meaningful differences in performance scores among the measured entities.

Our first strategy was to calculate common descriptive statistics that would help summarize the distribution of performance scores to see if there is meaningful variation across facilities. This included calculating the mean, median, standard deviation, and interquartile range of all the of the facility scores.

Our second strategy was to calculate a 95% confidence interval around each facility's score and to assess if the confidence interval included the national (or state) average. If the confidence interval did not include the national (or state) average, the facility was identified as being "better than average" or "worse than average". We also assessed if the lower bound of the 95% confidence interval was above 0.0, if so, this would indicate statistical confidence that misdiagnosis-related harms occurred.

**[Response Ends]**

**2b.06. Describe the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities.**

*Examples may include 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.*

**[Response Begins]**

***National hospital-level testing***

We plotted a histogram of the performance scores for the 967 facilities included in our sample.

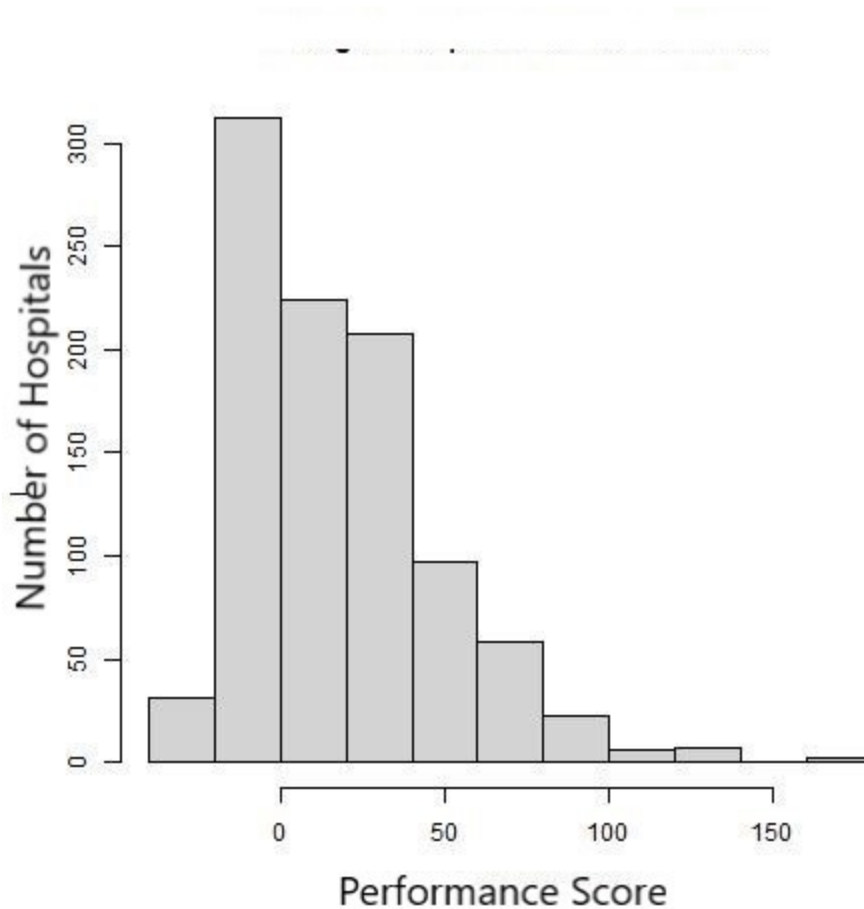


Figure 3. Histogram of Performance Scores for National Hospital-Level Testing

**Attributable 30-day Stroke Harms Rate (per 10,000 dizziness discharges)**

- Mean: 17.70
- Median: 13.33
- 25<sup>th</sup> Percentile: -7.32
- 75<sup>th</sup> Percentile: 31.43
- Standard Deviation: 30.04

**Better/Worse than National Average**

- 64.8% (n=627/967) hospitals were identified as being “better” than the national average (upper bound of 95% CI was less than national average)
- 0.8% (n=8/967) hospitals were identified as having statistically significant “harm” (lower bound of 95% CI was greater than zero)
- 0% (n=0/967) hospitals were identified as being “worse” than the national average (lower bound of 95% CI was greater than national average)

**State hospital-level testing**

We plotted a histogram of the performance scores for the 216 facilities included in our sample.

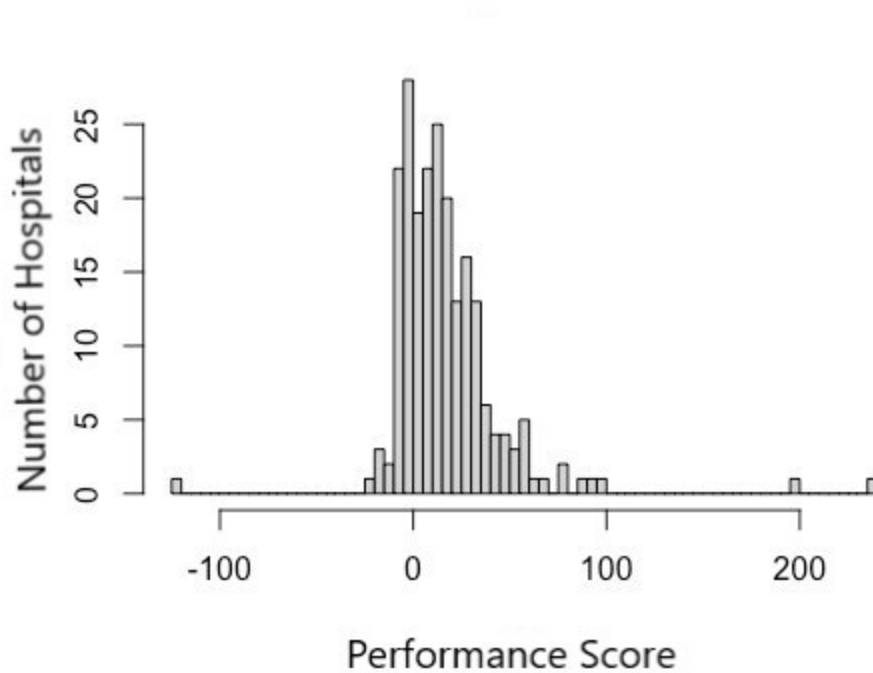


Figure 4. Histogram of Performance Scores for State Hospital-Level Testing

**Attributable 30-day Stroke Harms Rate (per 10,000 dizziness discharges)**

- Mean: 16.81
- Median: 11.27
- 25<sup>th</sup> Percentile: 0
- 75<sup>th</sup> Percentile: 26.92
- Standard Deviation: 29.86

**Better/Worse than State Average**

- 25.9% (n=56/216) hospitals were identified as being “better” than the state average (upper bound of 95% CI was less than state average)
- 6.5% (n=14/216) hospitals were identified as having statistically significant “harm” (lower bound of 95% CI was greater than zero)
- 0.9% (n=2/216) hospitals were identified as being “worse” than the state average (lower bound of 95% CI was greater than state average)

[Response Ends]

**2b.07. Provide 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.**

*In other words, what do the results mean in terms of statistical and meaningful differences?*

[Response Begins]

In both the national hospital-level and state hospital-level testing, we saw significant variation between facilities on the calculated measure with performance fairly evenly distributed around the median performance (i.e., difference between the median and 25<sup>th</sup> percentile is close to the difference between the median and the 75<sup>th</sup> percentile). Across the two datasets, the mean (17.7, 16.8 per 10,000) and median (13.3, 11.3 per 10,000) diagnostic performance measure scores were nearly identical.

With the measure, we were able to identify a sizable number of facilities who are “better than the national average.” But perhaps more importantly, we were able to identify a small number of facilities that had statistically significant rates of misdiagnosis “harm” or that were worse than the national or state averages.

The state hospital-level testing, which reflects effectively ~100% of measure-eligible ED “benign dizziness” discharges (rather than only the ~25% Medicare fraction available for the national hospital-level testing), demonstrates that the measure has even greater precision to identify differences among facilities when full data capture is possible.

As expected, the resolving power of the measure when using HCUP (state) dataset to determine “better or worse” hospitals was higher than that found when using the Medicare (national) data, since HCUP data include 100% of measure-eligible patient visits, while Medicare data include only ~25% of measure-eligible visits.

Facility-level diagnostic performance, when tested using either dataset, reveals mean and median performance of about 0.1-0.2% but with high outliers with missed stroke rates up to 1-2%—10-fold higher. For a medium- to large-sized hospital with 50,000 ED visits per year (~750 treat-and-release visits for “benign dizziness” each year, depending on patient mix) and an excess stroke hospitalization rate 10-fold over the mean at 1.7%, this would translate to 13 excess stroke hospitalizations after a misdiagnosis annually—more than one a month.

These results strongly argue that (a) the measure itself is precise enough to identify statistically significant and clinically meaningful differences across hospitals; (b) it is possible to identify data sources for benchmarking on this measure; and (c) it could be used to measure absolute harms, as well as both positive and negative deviance relative to the norm

**[Response Ends]**

**2b.08. Describe the method of testing conducted to identify the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders). Include how the specified handling of missing data minimizes bias.**

*Describe the steps—do not just name a method; what statistical analysis was used.*

**[Response Begins]**

#### ***National hospital-level testing***

Having access to the entire Medicare FFS dataset for our analysis provides us with one of the most comprehensive datasets available for quality measurement. The Medicare FFS data are already routinely used for calculating a large number of national performance measures for hospitals, including readmission rates and mortality rates. And while there may be a small number of Medicare beneficiaries that drop-out of FFS and then re-enter at a later point, we do not anticipate that the size of those numbers would be sizable enough to systematically bias our results.

#### ***State hospital-level testing***

From what we understand about the HCUP SEDD dataset for Florida covering the years 2016-2018, there is minimal, if any, missing data on the “benign dizziness” discharges from the ED, so we would not expect any bias in the denominator counts.

The potential for data missingness in a Florida-specific dataset is patients discharged from a Florida ED who are later admitted for stroke to a hospital outside of Florida. These stroke admissions would not be included in the Florida SID dataset.

As there is no systematic way for us to identify patients who were admitted to a hospital in another state for their stroke admission within the Florida SID dataset, we completed a number of sensitivity analyses to understand how a facility's performance on the measure could be impacted by a potential undercounting of stroke admissions. We discussed only adjusting the numerator counts for facilities that are located close to the state border (as determined by the predominant zip codes of patients who received care at the hospital), as these patients may be more likely to receive care in a neighboring state (Alabama, Georgia), but we finally decided that in a state like Florida, where there are many seasonal residents, adjusting just for facilities along the border may introduce its own bias. With input from subject matter experts on out-of-state hospital admissions, we concluded 5-10% of strokes being missed was a reasonable expectation of missingness.

For sensitivity analyses, we decided to re-calculate each facility's performance on the measure under the following scenarios:

<b>% increase in short-term strokes (1-30 days) to account for stroke admissions to hospitals outside of Florida</b>	<b>% increase in long-term strokes (91-360 days) to account for stroke admissions to hospitals outside of Florida</b>
5%	5%
5%	10%
10%	5%
10%	10%

Table 8. Scenarios Tested as Part of the Sensitivity Analyses for Missing Stroke Admissions (Short-Term and Long-Term)

Because the measure calculation incorporates both short-term strokes (likely "misdiagnosis") and longer-term strokes (baseline stroke rate), we thought it was important to consider the potential missingness in both of these counts. For example, with Florida having many seasonal residents during the Winter months, it is possible that some longer-term strokes are not captured in the SID dataset, as these patients may have returned to their primary home state 4-6 months after their ED visit.

**[Response Ends]**

**2b.09. Provide the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data.**

*For example, provide results of sensitivity analysis of the effect of various rules for missing data/non-response. If no empirical sensitivity analysis was conducted, identify the approaches for handling missing data that were considered and benefits and drawbacks of each).*

**[Response Begins]**

Below are how these four sensitivity analyses impact the statistically significant and clinically meaningful differences in facility performance on the measure, in comparison to the original calculations (n=216):

	Original calculations	5% short-term/5% long-term	5% short-term/10% long-term	10% short-term/5% long-term	10% short-term/10% long-term
Number of hospitals considered “Better than Average” (upper bound of 95% CI was less than state average)	14	17	15	20	20
Number of hospitals with statistically significant “harm” (lower bound of 95% CI was greater than zero)	56	57	57	58	58
Number of hospitals considered “Worse than Average” (lower bound of 95% CI was greater than state average)	2	2	2	2	2

Table 9. Results of the Sensitivity Analyses Scenarios for Missing Stroke Admissions (Short-Term and Long-Term).

**[Response Ends]**

**2b.10. Provide 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 non-responders), and how the specified handling of missing data minimizes bias.**

*In other words, 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 was conducted, justify the selected approach for missing data.*

**[Response Begins]**

As can be seen, there is very little difference in the estimates of overall hospital “better/worse” performance under any of these scenarios. This suggests that the results are likely robust to data missingness when using state-level data from HCUP. Within the small differences in the number of facilities classified as better/worse, there was slightly more potential impact on undercounting the number of facilities classified as “better than the state average,” with minimal impact on estimating “statistically significant harm” and no impact on classifying facilities as “worse than the state average.” While any misclassification is less than ideal, the misclassification does appear to minimize the potential for misclassification in ways that could impose “reputational harm” on a facility (i.e., being called “harmful” or “worse than average,” when actually not).

As previously mentioned, if full claims data capture were available for every hospital nationally, the missingness of the stroke admission data would be negligible. In other words, any potential problems with measure precision or accuracy linked to missingness would be fully mitigated by full access to appropriate data sets. In fact, the problem is principally one of data permissions—with access to fully de-identified Medicare and HCUP data, our results suggest that federal agencies such as CMS and AHRQ could readily benchmark across all institutions nationally with a high level of precision and accuracy. As described above, the two data resources (Medicare and HCUP) have complementary strengths and weaknesses that could be used to compensate for one other. For instance, Medicare data could be used to construct facility-specific “hospital crossover” or “state crossover” weights that could be applied to HCUP data to precisely and accurately benchmark performance for each hospital across the

nation. Alternatively, CMS could share a hospital-specific crossover weight with an individual hospital, which could then use their own data to calculate a crossover-weighted result with excellent precision and accuracy.

**[Response Ends]**

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 eQMs). 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.

**2b.11. Indicate whether there is more than one set of specifications for this measure.**

**[Response Begins]**

No, there is only one set of specifications for this measure

**[Response Ends]**

**2b.12. 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. Indicate what statistical analysis was used.*

**[Response Begins]**

**[Response Ends]**

**2b.13. Provide the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications.**

*Examples may include correlation, and/or rank order.*

**[Response Begins]**

**[Response Ends]**

**2b.14. Provide your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications.**

*In other words, what do the results mean and what are the norms for the test conducted.*

**[Response Begins]**

**[Response Ends]**

**2b.15. Indicate whether the measure uses exclusions.**

**[Response Begins]**

Yes, the measure uses exclusions.

**[Response Ends]**

**2b.16. Describe the method of testing exclusions and what was tested.**

*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?*

**[Response Begins]**

Not applicable

**[Response Ends]**

**2b.17. Provide 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.*

**[Response Begins]**

Not applicable

**[Response Ends]**

**2b.18. Provide your interpretation of the results, in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results.**

*In other words, 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.*

**[Response Begins]**

Not applicable

**[Response Ends]**

**2b.19. Check all methods used to address risk factors.**

**[Response Begins]**

No risk adjustment or risk stratification

Other approach to address risk factors (specify)

**[Response Ends]**

**2b.20. If using statistical risk models, provide detailed risk model specifications, including the risk model method, risk factors, risk factor data sources, coefficients, equations, codes with descriptors, and definitions.**

**[Response Begins]**

Not applicable

**[Response Ends]**

**2b.21. If an outcome or resource use measure is not risk-adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (i.e., case mix) is not needed to achieve fair comparisons across measured entities.**

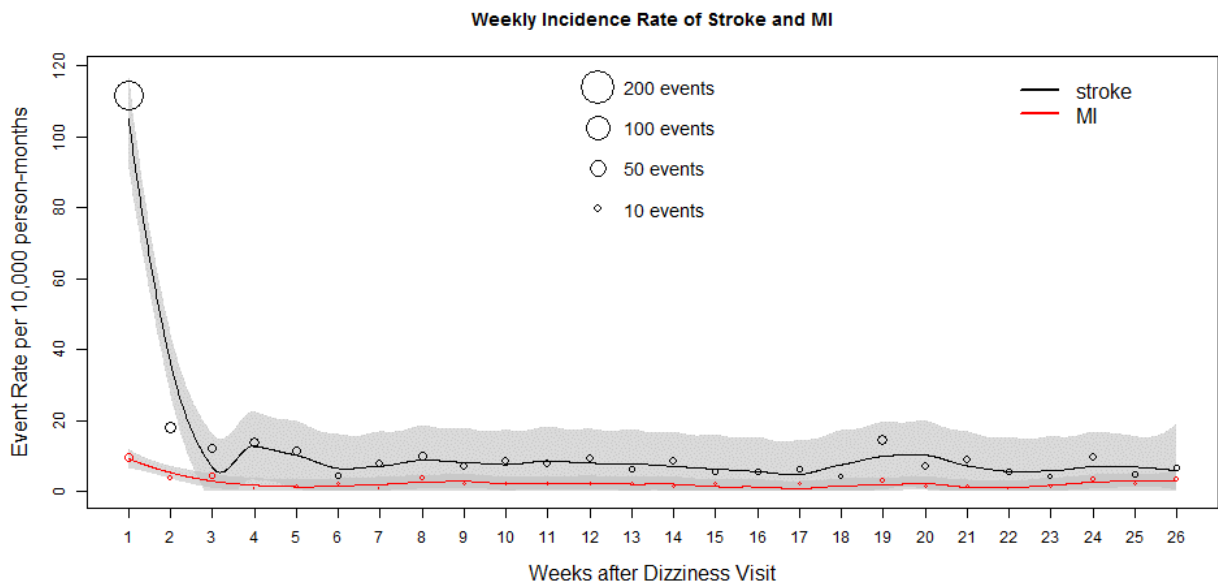
**[Response Begins]**

Our measure uses a statistical risk difference approach (observed [short-term stroke risk] minus expected [long-term/baseline stroke risk]) using the same patient cohort. As a result, controlling for differences in patients characteristics (case mix) is not needed to achieve fair comparisons across entities.

**Risk Difference Approach:** The risk-difference measure is a difference between two rates (observed minus expected), reflecting the observed stroke events in the first 30 days after an ED treat-and-release discharge (i.e., are likely to represent more than a chance association between the ED discharge and inpatient admission, above the expected epidemiologic base rate). This approach accounts for inter-institutional differences in the underlying stroke risk of their specific patient populations including any social determinants of long-term health in the affected population. It represents a conservative estimate of the rate of misdiagnosis-related harms from missed stroke because it assumes that long-term strokes (e.g., 91-360 days post discharge) are *not* likely to be preventable harms linked back to the original misdiagnosis.

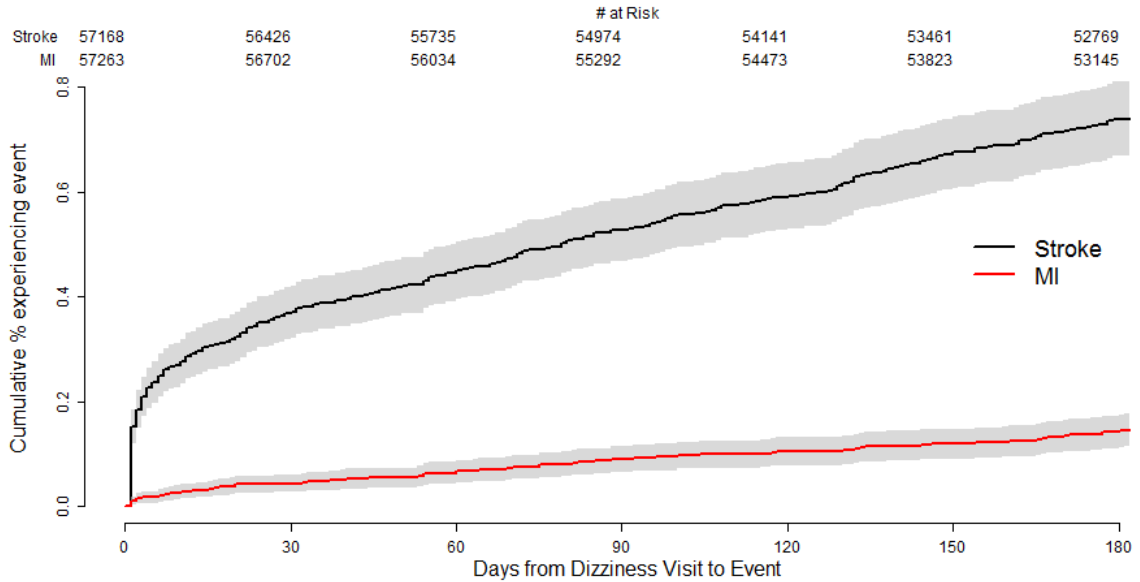
**Risk Difference Parameters:** The short-term **observed rate** is measured as the number of stroke hospitalizations per 10,000 discharges in the first 30 days and is called the **short-term 30-day rate of stroke hospitalization**. The short-term **expected rate** is estimated **in the exact same patients** by taking the average 30-day rate of stroke admission during a long-term outcome assessment window. The long-term window (91 days to 360 days post discharge) is chosen to reflect the epidemiologic base rate of stroke (i.e., after the short-term risk of a misdiagnosis leading to preventable major stroke has definitively passed). The stroke rate per 30-day period during this long-term 270-day window is obtained by dividing the numerator by nine and is called the **long-term 30-day rate**.

**Risk Difference Rationale:** Patients that have stroke hospitalizations within 30 days of an ED “benign dizziness” discharge represent patients that are misdiagnosed at the ED index visit, but also include some patients that are not misdiagnosed (i.e., do, in fact, have benign dizziness) who go on have a coincidental stroke event due to baseline (biological/sociocultural) stroke risk. This baseline stroke risk is reflected by the long-term population-specific stroke rate which is not related to the institutional rate of misdiagnosis or short-term harms (i.e., 30-day stroke admissions). This relationship is most evident when viewed as a longitudinal incidence rate curve for stroke hospitalization (Fig. 5). This curve matches the natural history/biological profile of major stroke following minor stroke and TIA (Fig. 6/7).

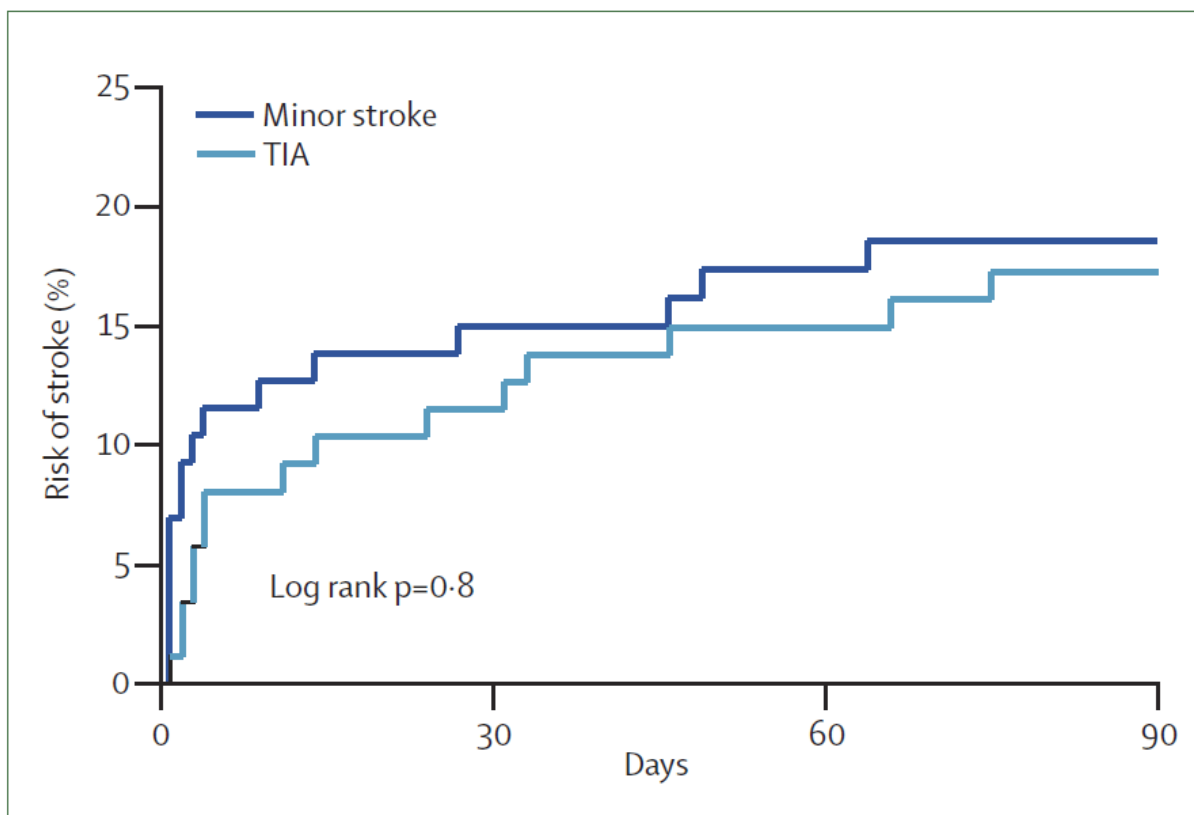


**Figure 5. Weekly incidence rate curve of stroke hospitalizations post ED treat-and-release discharge as “benign dizziness.”**

Kaiser Permanente Mid-Atlantic data from the performance period from 2010-2014 at all outpatient sites (ED, ambulatory care). Data reflect 56,746 treat-and-release visits for “benign dizziness.” Shown in black are stroke hospitalizations, and shown in red are heart attack hospitalizations (for comparison). Gray shading represents 95% confidence intervals for each. Early returns for stroke hospitalization above the epidemiologic base rate in the first few weeks after discharge reflect potentially preventable harms from stroke missed at the index visit. The comparison outcome of heart attack demonstrates the association is specific for dizziness and stroke (i.e., absent for dizziness and hear attack).



**Figure 6. Cumulative incidence curve of stroke hospitalizations post ED treat-and-release discharge as “benign dizziness.”** Represented here are the same data as shown in Figure 5. These data are presented here as a cumulative incidence curve for comparison to Figure 7, which illustrates the disease natural history of major stroke after transient ischemic attack (TIA) or minor stroke. Note the similarities between the top-most curves shown in Figures 6 (stroke hospitalizations after a “dizziness” discharge) and 7 (major stroke after minor stroke).



**Figure 7. Cumulative incidence curve for major stroke following TIA or minor stroke.** Data are from the Oxford Vascular Study as represented in Rothwell, Buchan, & Johnston (*Lancet Neurol*, 2006; PMID: 16545749). This natural history curve matches the empirical pattern of stroke hospitalizations when some patients are diagnosed (erroneously) as “benign dizziness” and discharged home (Figure 6).

This risk difference approach uses an institution-specific longer-term (91d–360d) stroke hospitalization rate to approximate the baseline short-term stroke risk for the population in question. This long-term window is chosen because, biologically speaking, the short-term risk of major stroke after minor stroke or TIA levels off by approximately 30 days after the initial cerebrovascular event (Figure 7). By using the risk difference, the measure quantifies only the “excess” short-term stroke rate (attributable risk) due to misdiagnosis above the base rate for the population in question. Thus, the risk difference accounts for all relevant demographic differences across populations including biological and social and determinants of health that may lead to population-level variation in baseline stroke risk.

**Rationale for No Demographic Risk Adjustments:** Other racial or demographic disparities in institution-specific risk of misdiagnosis that are linked to the institution-specific patient populations should be measured appropriately rather than “adjusted” away (e.g., racial bias, racial minorities are at higher risk of being misdiagnosed [Newman-Toker et al., *Diagnosis*, 2014; PMID: 28344918]).

**Risk Difference Calculation:** The risk difference calculation requires an observed and expected rate calculation. For each patient discharged from the ED with a “benign dizziness” diagnosis during the performance period, data on stroke hospitalizations must be available for a floating outcome assessment window of roughly 12 months (360 days). If stroke hospitalizations occur between post-ED day #1 and day #30 (i.e., mostly linked to misdiagnosis-related harms), they are counted in the numerator of the “short-term 30-day rate” (observed rate). If stroke hospitalizations occur between post-ED day #91 and day #360 (i.e., mostly linked to baseline biological or sociocultural stroke risk), they are counted in the numerator of the “long-term 30-day rate.” The long-term rate is normalized to a 30-day period equivalent rate over the 270-day outcome assessment window by dividing by nine

(i.e., taking the average 30-day rate during those 270 days). A 270-day window is used for the average longer-term 30-d rate calculation because of very low stroke base rates in this time window (<0.1% [Newman-Toker, *Ann Neurol*, 2015; PMID: 26418192]); this increases the precision of the “expected” value.

- **Crude short-term 30-day rate** = {[number of stroke hospitalizations within 30d + *alpha*] / [number of eligible ED benign dizziness discharges in the performance period + 1]} x 10,000. This “short-term” rate includes the early peak rate (Fig. 1) of hospitalization after missed stroke and dominantly reflects misdiagnosis (but partly reflects the base rate). The measure is represented as the number of stroke hospitalizations per 10,000 benign dizziness discharges. The constants “*alpha*” = 1/1,000 and “1” are added to avoid issues with possible zero counts.
- **Crude long-term 30-day rate** = {[number of stroke hospitalizations from 91d-360d divided by 9] + *alpha*] / [number of eligible ED benign dizziness discharges in the performance period and no stroke diagnosis in the prior 90 days + 1 - (3 x *alpha*)]} x 10,000. This “long-term” rate approximates the epidemiologic “base” rate of stroke in the specific population in whom the short-term 30d rate is measured. The parameter is represented as the number of stroke hospitalizations per 10,000 benign dizziness discharges. The denominator should exclude those patients who experienced a stroke prior to 90 days since we are only counting the first stroke in the 360 days post index visit. The constants “*alpha*” = 1/1,000 and “1 - [3 x *alpha*]” are added to avoid issues with possible zero counts.
- **Attributable short-term 30d rate** = (crude short-term 30d rate) – (crude long-term 30d rate); the attributable short-term rate reflects the “excess” short-term (30d) rate of stroke above the base rate that is specific for the population in question. This is an estimate of the **attributable risk** of misdiagnosis-related harms from missed stroke. The parameter is represented as the number of stroke hospitalizations per 10,000 benign dizziness discharges.

[Response Ends]

**2b.22. Select all applicable resources and methods used to develop the conceptual model of how social risk impacts this outcome.**

[Response Begins]

Other (specify)

[Other (specify) Please Explain]

Not applicable

[Response Ends]

**2b.23. Describe the conceptual and statistical methods and criteria used to test and select patient-level risk factors (e.g., clinical factors, social risk factors) used in the statistical risk model or for stratification by risk.**

*Please be sure to address the following: potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of  $p < 0.10$  or other statistical tests; correlation of  $x$  or higher. Patient factors should be present at the start of care, if applicable. Also discuss any “ordering” of risk factor inclusion; note whether social risk factors are added after all clinical factors. Discuss any considerations regarding data sources (e.g., availability, specificity).*

[Response Begins]

Removing the expected rate based on the same cohort accounts for all relevant clinical and social risk factors that contribute to baseline biologic risk of subsequent major stroke after minor stroke or TIA. Thus, there was no need to assign or measure specific patient factors in this calculation.

No clinical or social risk factors are used to adjust the observed rate. This is because demographic disparities in institution-specific risk of misdiagnosis that are linked to the institution-specific patient population should be measured appropriately rather than “adjusted” away (e.g., racial bias which may place minorities at higher risk of being misdiagnosed [PMID: 28344918])

[Response Ends]

**2b.24. Detail the statistical results of the analyses used to test and select risk factors for inclusion in or exclusion from the risk model/stratification.**

[Response Begins]

Not applicable

[Response Ends]

**2b.25. Describe the analyses and interpretation resulting in the decision to select or not select social risk factors.**

*Examples may include prevalence of the factor across measured entities, availability of the data source, empirical association with the outcome, contribution of unique variation in the outcome, or assessment of between-unit effects and within-unit effects. Also describe the impact of adjusting for risk (or making no adjustment) on providers at high or low extremes of risk.*

[Response Begins]

Not applicable

[Response Ends]

**2b.26. 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 control for differences in patient characteristics (i.e., case mix) below. If stratified ONLY, enter “N/A” for questions about the statistical risk model discrimination and calibration statistics.**

*Validation testing should be conducted in a data set that is separate from the one used to develop the model.*

[Response Begins]

Not applicable

[Response Ends]

**2b.27. Provide risk model discrimination statistics.**

*For example, provide c-statistics or R-squared values.*

[Response Begins]

Not applicable

[Response Ends]

**2b.28. Provide the statistical risk model calibration statistics (e.g., Hosmer-Lemeshow statistic).**

**[Response Begins]**

Not applicable

**[Response Ends]**

**2b.29. Provide the risk decile plots or calibration curves used in calibrating the statistical risk model.**

*The preferred file format is .png, but most image formats are acceptable.*

**[Response Begins]**

Not applicable

**[Response Ends]**

**2b.30. Provide the results of the risk stratification analysis.**

**[Response Begins]**

Not applicable

**[Response Ends]**

**2b.31. Provide your interpretation of the results, in terms of demonstrating adequacy of controlling for differences in patient characteristics (i.e., case mix).**

*In other words, what do the results mean and what are the norms for the test conducted?*

**[Response Begins]**

Not applicable

**[Response Ends]**

**2b.32. Describe any additional testing conducted to justify the risk adjustment approach used in specifying the measure.**

*Not required but would provide additional support of adequacy of the risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed.*

**[Response Begins]**

Not applicable

**[Response Ends]**

### 3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

---

**3.01. Check all methods below that are used to generate the data elements needed to compute the measure score.**

[Response Begins]

[Response Ends]

**3.02. Detail to what extent the specified data elements are available electronically in defined fields.**

*In other words, indicate whether data elements that are needed to compute the performance measure score are in defined, computer-readable fields.*

[Response Begins]

[Response Ends]

**3.03. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using data elements not from electronic sources.**

[Response Begins]

[Response Ends]

**3.04. Describe any efforts to develop an eCQM.**

[Response Begins]

[Response Ends]

**3.06. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

[Response Begins]

[Response Ends]

Consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

**3.07. Detail any fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, algorithm),**

**Attach the fee schedule here, if applicable.**

[Response Begins]

[Response Ends]



## 4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

---

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement, in addition to demonstrating performance improvement.

**4a.01. Check all current uses. For each current use checked, please provide:**

- **Name of program and sponsor**
- **URL**
- **Purpose**
- **Geographic area and number and percentage of accountable entities and patients included**
- **Level of measurement and setting**

[Response Begins]

[Response Ends]

**4a.02. Check all planned uses.**

[Response Begins]

[Response Ends]

**4a.03. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing), explain why the measure is not in use.**

*For example, do policies or actions of the developer/steward or accountable entities restrict access to performance results or block implementation?*

[Response Begins]

[Response Ends]

**4a.04. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes: used in any accountability application within 3 years, and publicly reported within 6 years of initial endorsement.**

*A credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*

[Response Begins]

[Response Ends]

**4a.05. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**

*Detail how many and which types of measured entities and/or others were included. If only a sample of measured entities were included, describe the full population and how the sample was selected.*

[Response Begins]

[Response Ends]

**4a.06. Describe the process for providing measure results, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**

[Response Begins]

[Response Ends]

**4a.07. Summarize the feedback on measure performance and implementation from the measured entities and others. Describe how feedback was obtained.**

[Response Begins]

[Response Ends]

**4a.08. Summarize the feedback obtained from those being measured.**

[Response Begins]

[Response Ends]

**4a.09. Summarize the feedback obtained from other users.**

[Response Begins]

[Response Ends]

**4a.10. Describe how the feedback described has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**

[Response Begins]

[Response Ends]

**4b.01. You may refer to data provided in Importance to Measure and Report: Gap in Care/Disparities, but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included). If no improvement was demonstrated, provide an explanation. If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

[Response Begins]

[Response Ends]

**4b.02. Explain any unexpected findings (positive or negative) during implementation of this measure, including unintended impacts on patients.**

**[Response Begins]**

**[Response Ends]**

**4b.03. Explain any unexpected benefits realized from implementation of this measure.**

**[Response Begins]**

**[Response Ends]**

## 5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

---

If you are updating a maintenance measure submission for the first time in MIMS, please note that the previous related and competing data appearing in question 5.03 may need to be entered in to 5.01 and 5.02, if the measures are NQF endorsed. Please review and update questions 5.01, 5.02, and 5.03 accordingly.

### 5.01. Search and select all NQF-endorsed related measures (conceptually, either same measure focus or target population).

**NOTE: If there are no related measures, please select N/A.**

*(Can search and select measures.)*

[Response Begins]

[Response Ends]

### 5.02. Search and select all NQF-endorsed competing measures (conceptually, the measures have both the same measure focus and target population).

**NOTE: If there are no competing measures, please select N/A.**

*(Can search and select measures.)*

[Response Begins]

[Response Ends]

### 5.03. If there are related or competing measures to this measure, but they are not NQF-endorsed, please indicate the measure title and steward.

[Response Begins]

[Response Ends]

### 5.04. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s), indicate whether the measure specifications are harmonized to the extent possible.

[Response Begins]

[Response Ends]

### 5.05. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

[Response Begins]

[Response Ends]

### 5.06. Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality). Alternatively, justify endorsing an additional measure.

*Provide analyses when possible.*

[Response Begins]

[Response Ends]

## Appendix

Supplemental materials may be provided in an appendix.:

## Contact Information

**Measure Steward (Intellectual Property Owner):** Johns Hopkins Armstrong Institute for Patient Safety and Quality

**Measure Steward Point of Contact:** Newman-Toker, David, [toker@jhu.edu](mailto:toker@jhu.edu)

Austin, Matt, [jausti17@jhmi.edu](mailto:jausti17@jhmi.edu)

**Measure Developer if different from Measure Steward:** Johns Hopkins Armstrong Institute for Patient Safety and Quality

**Measure Developer Point(s) of Contact:** Newman-Toker, David, [toker@jhu.edu](mailto:toker@jhu.edu)

Austin, Matt, [jausti17@jhmi.edu](mailto:jausti17@jhmi.edu)

## Additional Information

1. Provide any supplemental materials, if needed, as an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be collated one file with a table of contents or bookmarks. If material pertains to a specific criterion, that should be indicated.

[Response Begins]

[Response Ends]

2. List the workgroup/panel members' names and organizations.

*Describe the members' role in measure development.*

[Response Begins]

[Response Ends]

3. Indicate the year the measure was first released.

[Response Begins]

[Response Ends]

4. Indicate the month and year of the most recent revision.

[Response Begins]

[Response Ends]

5. Indicate the frequency of review, or an update schedule, for this measure.

[Response Begins]

[Response Ends]

6. Indicate the next scheduled update or review of this measure.

[Response Begins]

[Response Ends]

7. Provide a copyright statement, if applicable. Otherwise, indicate "N/A".

[Response Begins]

[Response Ends]

8. State any disclaimers, if applicable. Otherwise, indicate "N/A".

[Response Begins]

[Response Ends]

9. Provide any additional information or comments, if applicable. Otherwise, indicate "N/A".

[Response Begins]

**[Response Ends]**