# Measure Worksheet (MEW-PA-New)

Partnership for Quality Measurement

Click here for Pre-Evaluation Public Comments

Click here for Measure Specifications

## Content

### Brief Measure Information

**CBE #:** 3746

**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 incidence rate, reflecting days 0-30 days] minus expected [long-term incidence rate, reflecting days 91-360]).

**1b.01. Developer Rationale:** Diagnostic error is a major public health problem.40 The lack of operational measures is a critical barrier to improving diagnostic quality.41,42 Three major disease categories (vascular events, infections, and cancer) account for three-fourths of all serious harms from diagnostic error as identified by malpractice claims.43 Among vascular events, missed stroke is the leading cause of serious harm to patients. Misdiagnosis of stroke disproportionately occurs when patients present with symptoms/signs that are not typical or obvious for stroke.8,44 For example, the most common clinical presentation of missed stroke occurs when patients present with dizziness or vertigo, which can easily be mistaken for inner ear disease.8 Annually in US emergency departments (ED), an estimated 45,000-75,000 patients that present with dizziness or vertigo and have strokes, are misdiagnosed and erroneously discharged from the ED.44

ED patients with acute dizziness and vertigo could be correctly diagnosed with stroke using evidence-based bedside examinations,3,35 but there is a large evidence-practice gap38 in ED diagnosis, resulting in substantial harms to patients.44 Without a timely and accurate diagnosis, these patients suffer misdiagnosis-related harms45 because they do not receive prompt treatment for this time-sensitive condition.8 The most common harm is a preventable major stroke leading to a subsequent hospitalization after the patient has had a minor stroke or transient ischemic attack (TIA).21,22 Crude short-term stroke hospitalization rates per 10,000 ED dizziness discharges vary at least from 20-80.44 Adjusting for baseline stroke risk across groups does not eliminate practice variation.46

This outcome measure tracks the rate of missed strokes in the ED—i.e., patients admitted to the hospital for a stroke within 30 days of an ED discharge with a non-specific diagnosis of benign dizziness diagnosis or a specific inner ear/vestibular diagnosis

| Content |
| --- |
| (collectively referred to as “benign dizziness”). This measure is the first operationally viable performance measure of stroke misdiagnosis in the hospital setting. Hospital EDs will be able to use the measure to internally track their performance over time as they work to implement interventions to reduce stroke misdiagnosis. The measure can also be used by external entities for public reporting and pay-for-performance, as external pressure to encourage improvement in diagnostic quality. |
| **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 hospital discharge diagnosis of stroke.<br><br>**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.”<br><br>**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<br><br>**sp.28. Data Source:** Claims<br><br>**sp.07. Level of Analysis:** Facility |
| **IF Endorsement Maintenance—Original Endorsement Date:** N/A New measure<br><br>**Most Recent Endorsement Date:** N/A New measure |
| **IF this measure is included in a composite, Composite#/title:** N/A<br><br>**IF this measure is paired/grouped, CBE#/title:** N/A<br><br>**sp.03. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** |
| **Staff Assessment: New Measure** |
| **Criterion 1: Importance to Measure and Report** |
| **1a. Evidence** |

**Content**

**1a. Evidence.** The evidence requirements for a ***health outcome*** measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance can be used, assuming the data are from a robust number of providers and the results are not subject to systematic bias. For measures derived from a patient report, the evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

**The developer provides the following description for this measure:**

- This is a new outcome measure at the facility-level that captures 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 hospital discharge diagnosis of stroke.
- The developer provides a logic model that depicting that improvement on this measure requires quality improvement (QI) efforts that improve diagnosis of ED patients that present with dizziness/vertigo. The developer states that these QI efforts will improve diagnosis both for patients with stroke and patients with inner ear disease. Benefits to patients will accrue from the prompt application of research-proven treatments:
  - Those with stroke will benefit from tPA or early secondary prevention, as appropriate
  - Those with benign paroxysmal positional vertigo will benefit from prompt canalith repositioning and less CT radiation
- These benefits to stroke patients, in turn, will, as the developer notes, result in a "better" measure score.

**Summary:**

- The devleoper cites several studies, including systematic reviews, supporting the notion that dizziness is frequently misdiagnosed in the ED, and that better medical care (i.e., better neurological examinations) may result in fewer misdiagnoses.
- The developer also cites evidence of meaningfulness to patients. For patients seeking care in the ED with new dizziness, the devleoper notes that worry about the cause can be prominent, often including fear of having a stroke.
- The developer further underscores that clinicians have "wanted a valid approach to 'help decide whether to obtain neuroimaging' or 'exclude stroke as a cause of dizziness in ED patients without neuroimaging.'"
- By conducting an internal analysis using the full national Medicare fee-for-service dataset, the devleoper explored the relationship between a hospital's performance on the dizzy-stroke measure and its use of imaging in dizzy patients and the type of imaging used (CT vs. MRI). The results show that a higher imaging rate of any type is associated with a lower rate of misdiagnosis, particularly if the imaging is by MRI (Table 2). Imaging by CT is linked to an increased risk of adverse outcomes from missed stroke; this effect likely represents correct clinical risk stratification and false reassurance by falsely negative CT neuroimaging for acute ischemic stroke, common in dizziness. The developer notes that these findings suggest "that one possible intervention to reduce missed strokes is to obtain more MRIs and fewer CTs in appropriate patients with dizziness or vertigo."

***Question for the Standing Committee:***

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

**Guidance From the Evidence Algorithm**

**Content**

Box 1: Yes → Box 2: Pass

**Preliminary rating for evidence:** ☒ **Pass** ☐ **No Pass**

**1b. Gap in Care/Opportunity for Improvement and Disparities**

**1b. Performance Gap.** The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer reports performance scores from 1/1/2015 – 12/31/2017:
  - Data Source: Medicare Fee-for-Service + Medicare Advantage
  - Number of Measured Entities: 967 Hospital Eds
  - Number of Patients: 383,017
  - Mean Score: 17.70; SD: 30.04;
  - Min Score: (-29.15); Max Score: 165.32; IQ Range: (-7.32, 31.43);
  - Median scores by decile: (-17.58, -12.10, -7.35, 0.00, 10.41, 16.91, 23.54, 31.44, 49.62, 73.66)
- The developer also provides similar results from prior years ([1/1/2012-12/31/2014] and [1/1/2009-12/31/2011])

**Disparities**

- The developer cites evidence that women and minorities are at ~20-30% increased odds of stroke misdiagnosis and patients 18-44 years old are at roughly 7-fold increased odds.

***Questions for the Standing Committee:***

- *Is there a gap in care that warrants a national performance measure?*

**Preliminary rating for opportunity for improvement:**
☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

**Criteria 2: Scientific Acceptability of Measure Properties**

**Complex measure evaluated by the Scientific Methods Panel (SMP)?** ☐ **Yes** ☒ **No**

**Evaluators:** Battelle Staff

**2a. Reliability: Specifications and Testing**

**2a2. Reliability testing** demonstrates whether the measure data elements are repeatable and producing the same results a high proportion of the time when assessed in the same population in the same time period, and/or whether the measure score is precise enough to distinguish differences in performance across providers.

**Specifications:**

- Measure specifications are clear and precise.

**Reliability Testing:**

- Reliability testing conducted at the Accountable Entity Level:
  - The developer conducted a signal-to-noise analysis at both the national hospital-level (using Medicare FFS and

| Content |
| --- |
| Medicare Advantage data) and state hospital-level (using Florida HCUP data) from 1/1/2015 – 12/31/2017 and 1/1/2016 – 12/31/2017, respectively.<br>○ For the national hospital-level:<br>○ The median reliability score for the entire 967-hospital sample was 0.590, with an interquartile range of 0.414-0.951.<br>○ In a stratified analysis, the developer reports that reliability increases when the number of visits analyzed increases.<br>○ For the state hospital-level:<br>○ The median reliability score for the entire 216-hospital sample was 0.853, with an interquartile range of 0.671-0.950.<br>○ The developer posits that reliability was much higher in the state-level analysis than in the national-level analysis because of data missingness in Medicare data (i.e., Medicare represents only ~25% of eligible ED index visits, largely because of the age constraint [mostly patients ≥65yo]). |
| ***Questions for the Standing Committee regarding reliability:***<br>• *Do you have any concerns that the measure cannot be consistently implemented (i.e., are the measure specifications adequate)?* |
| **Guidance From the Reliability Algorithm**<br>Box 1: Yes → Box 2: Yes → Box 4: Yes → Box 5: Yes → Box 6: Moderate<br><br>The highest possible rating is HIGH. |
| **Preliminary rating for reliability:** ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient** |
| **2b. Validity:** **Validity Testing; Exclusions; Risk Adjustment; Meaningful Differences; Comparability; Missing Data** |
| **2b2. Validity testing** should demonstrate that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. |
| **2b2-2b6. Potential threats to validity** should be assessed/addressed.<br><br>**Validity Testing**<br>• Validity testing conducted at the Patient/Encounter Level:<br>○ For data element validity for stroke, the developers cited prior literature that used claims data to identify stroke discharges using chart abstraction as the standard (numerator).<br>○ For denominator for benign dizziness diagnoses, the developer conducted two studies focused on code-level validity. First, when an ED patient has a “benign dizziness” discharge diagnosis, how often do charts suggest the ED provider intended to code “benign dizziness”? This was conducted using two academic hospitals. PPV was calculated in a random sample of 64 charts in three cohorts (i.e., chief complaints of dizziness, oto-vestibular complaints, and other chief<br>○ complaints). Second, the developer calculated an NPV specifically if another diagnosis was coded; how often did they |

| Content |
| --- |
| intend to code something other than benign dizziness? They reviewed a random sub-sample of 67 charts for high-risk sub-group to estimate NPV. The PPV was 100 percent for coding benign dizziness. The NPV was nearly 100 percent. The audit of discharged status demonstrated 100 percent accuracy, even for the highest risk cases. |
| **Exclusions**<br>• The measure does not use exclusions. |
| **Risk Adjustment**<br>• The measure is not risk-adjusted or stratified.<br>• However, the developer states that the 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 patient characteristics (case mix) is not needed to achieve fair comparisons across entities.<br>• The devleoper adds that 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. |
| **Meaningful Differences**<br>• The devleoper provided a distribution of measure scores at the national and state hospital-levels.<br>• At the national hospital-level:<br>  • Attributable 30-day Stroke Harms Rate (per 10,000 dizziness discharges)<br>  • Mean: 17.70<br>  • Median: 13.33<br>  • 25th Percentile: -7.32<br>  • 75th Percentile: 31.43<br>  • Standard Deviation: 30.04<br>  • Better/Worse than National Average<br>  • 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)<br>  • 0.8% (n=8/967) hospitals were identified as having statistically significant "harm" (lower bound of 95% CI was greater than zero)<br>  • 0% (n=0/967) hospitals were identified as being "worse" than the national average (lower bound of 95% CI was greater than national average)<br>• At the state hospital-level:<br>  • Attributable 30-day Stroke Harms Rate (per 10,000 dizziness discharges)<br>  • Mean: 16.81<br>  • Median: 11.27<br>  • 25th Percentile: 0<br>  • 75th Percentile: 26.92<br>  • Standard Deviation: 29.86<br>  • Better/Worse than State Average<br>  • 25.9% (n=56/216) hospitals were identified as being "better" than the state average (upper bound of 95% CI was less |

| Content |
| --- |
| than state average)<br>• 6.5% (n=14/216) hospitals were identified as having statistically significant “harm” (lower bound of 95% CI was greater than zero)<br>• 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) |
| **Missing Data**<br>• With respect to the national Medicare FFS dataset, the developer states that 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, the developer does not anticipate that the size of those numbers would be sizable enough to systematically bias our results.<br>• For the state hospital-level, the developer noted that 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.<br>• The developer further states that there is no systematic way to identify patients who were admitted to a hospital in another state for their stroke admission within the Florida SID dataset. Therefore, the developer 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.<br>• The developer reported that there is very little difference in the estimates of overall hospital “better/worse” performance, suggesting that the results are likely robust to data missingness when using state-level data from HCUP. |
| **Comparability**<br>• The measure only uses one set of specifications for this measure. |
| ***Questions for the Standing Committee regarding validity:***<br>• *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk adjustment approach, etc.)?* |
| **Guidance From the Validity Algorithm**<br>Box 1: Yes → Box 2: Yes → Box 5: No → Box 9: Yes → Box 10: Yes → Box 11: Moderate<br><br>The highest possible rating is MODERATE |
| **Preliminary rating for validity:** ☒ **Moderate** ☐ **Low** ☐ **Insufficient** |
| **Criterion 3. Feasibility** |
| **3.** Feasibility is the 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.<br>• The developer notes that the measure requires very few data elements in order to be calculated, all of which are routinely collected in the course of clinical care – discharge diagnosis codes (ICD-10-CM) and dates for emergency department (ED) visits and inpatient hospital stays.<br>• All data elements are in defined fields in electronic claims. |

**Content**

- The developer adds that there are no explicit fees or licenses associated with calculating this measure.

***Questions for the Standing Committee:***

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form (e.g., EHR or other electronic sources)?*

**Preliminary rating for feasibility: ☒ High ☐ Moderate ☐ Low ☐ Insufficient**

**Criterion 4: Use and Usability**

**4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)**

**4a. Use** evaluates the extent to which audiences (e.g., consumers, purchasers, providers, and policymakers) use or could use performance results for both accountability and performance improvement activities.

**4a1. Accountability and Transparency.** Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If they are not in use at the time of initial endorsement, then a credible plan for implementation within the specified time frames is provided.

**Current uses of the measure**

| | | | |
|---|---|---|---|
| Publicly reported? | ☐ Yes | ☒ No | |
| Current use in an accountability program? | ☐ Yes | ☒ No | ☐ UNCLEAR |
| Planned use in an accountability program? | ☒ Yes | ☐ No | ☐ N/A |

**Accountability program details**

- This new measure is currently not currently publicly reported or used within an accountability application. However, the developer notes that the measure is being reported to ED quality and safety leaders and the Director of the Armstrong Institute for Patient Safety and Quality at Johns Hopkins (who is also Sr. VP, Patient Safety and Quality for Johns Hopkins Medicine) on an annual basis, as recommended for the current measure parameterization (3-year rolling window updated annually).
- However, the developer states that it can be used for internal quality improvement within hospitals and that the measure lends itself to having a federal agency, such as the Agency for Healthcare Research and Quality (AHRQ), calculate aggregated hospital performance using a national dataset (e.g., HCUP dataset) and track national performance on the measure over time.
- For public reporting and use within a payment program, the developer suggests that public reporting and external benchmarking initially on a voluntary basis could occur through the Leapfrog Group and that it anticipates that the measure could be incorporated into hospital pay-for-performance programs, with possible adoption by the Centers for Medicare and Medicaid Services (CMS) and other payers.

**4a.2. Feedback on the measure by those being measured or others.** Three criteria demonstrate feedback: (1) Those being

| Content |
| --- |
| measured have been given performance results or data, as well as assistance with interpreting the measure results and data; (2) Those being measured, and other users have been given an opportunity to provide feedback on the measure performance or implementation; and (3) This feedback has been considered when changes are incorporated into the measure. |
| **Feedback on the measure provided by those being measured or others**<br>• As mentioned above, the measure is being reported to ED quality and safety leaders and the Director of the Armstrong Institute for Patient Safety and Quality at Johns Hopkins (who is also Sr. VP, Patient Safety and Quality for Johns Hopkins Medicine) on an annual basis, as recommended for the current measure parameterization (3-year rolling window updated annually).<br>• Due to this use, the developer attests that feedback on the measure from ED physicians in the quality improvement space has been very positive, overall.<br>• The developer notes that feedback has led to modified use of code sets for the stroke numerator. On the basis of feedback, a modified denominator version (using a presenting symptom of dizziness, rather than a discharge diagnosis), is being developed in parallel. |
| ***Questions for the Standing Committee:***<br>• *Can the performance results be used to further the goal of high quality, efficient healthcare?*<br>• *How has the measure been vetted in real-world settings by those being measured or others?* |
| **Preliminary rating for Use:** ☒ **Pass** ☐ **No Pass** |
| **4b. Usability (4b1. Improvement; 4b2. Benefits of measure)** |
| **4b. Usability** evaluates the extent to which audiences (e.g., consumers, purchasers, providers, and policymakers) use or could use performance results for both accountability and performance improvement activities. |
| **4b1 Improvement.** Progress toward achieving the goal of high quality, efficient healthcare for individuals or populations is demonstrated.<br><br>**Improvement results**<br>• The developer reports that across the three 3-year time periods for which the measure was calculated, there has been a small, but steady improvement over time.<br>• The mean performance on the measure has improved slightly in each successive time period (where lower performance is desirable) and the standard deviation on the measure has shrunk. Despite this apparent improvement, the devleoper adds that the median hospital performance on the measure in 7 of the 10 deciles remains at or above zero, indicating there is still room for improvement at most hospitals. |
| **4b2. Benefits versus harms.** The benefits of the performance measure in facilitating progress toward achieving high quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists). |

| Content |
| --- |
| **Unexpected findings (positive or negative) during implementation**<br>• The developer attests that it has detected no unexpected findings (positive or negative) during the relatively recent and small-scale deployment of this measure, including no unintended impacts on patients. |
| ***Questions for the Standing Committee:***<br>• *How can the performance results be used to further the goal of high quality, efficient healthcare?*<br>• *Do the benefits of the measure outweigh any potential unintended consequences?* |
| **Preliminary rating for Usability and Use:**<br>☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient** |
| **Criterion 5: Related and Competing Measures** |
| **Related Measures**<br>• The developer reports that no related or competing measures have been identified. |
| **Harmonization**<br>• N/A |

**QUALITY MEASURE SUBMISSION FORM**

**Version: 1.0; Generated: 13 April 2023**

## Introduction

Thank you for your interest in submitting a measure to Battelle for possible endorsement.

**What criteria are used to evaluate measures?** Measures are evaluated on standardized criteria: importance to measure and report, scientific acceptability of measure properties, feasibility, usability and use, and related and competing measures. For your measure to be evaluated against these measure evaluation criteria, you must complete the measure submission form.

**Why do I have to complete a form?** Due to the volume and/or complexity of proposed measures, Battelle provides measure information to committee reviewers in a standardized format to facilitate their evaluation of whether the measure meets the measure evaluation criteria. This for allows the measure steward to present information demonstrating that the proposed measure meets endorsement criteria.

**What is on the form?** The information requested in this form is directly related to the measure evaluation criteria.

**Can't I just submit our files for consideration?** No. Measures must be submitted through the online form to be considered for the Spring 2023 cycle. Requested information should be entered directly into this form and as well as any necessary or required attachments.

**Can I submit additional details and materials?** Additional materials will be considered only as supplemental. Do NOT rely on material provided in an appendix to provide measure specifications or to demonstrate meeting the criteria. The core information needed to evaluate the measure should be provided in the appropriate submission form fields and required attachments. Please contact PQMsupport@battelle.org regarding questions about submitting supplemental materials.

**What do I do first?** If you have started a new submission by answering five qualifying questions, you may proceed to the “Previous Submission Information" tab to continue with your submission. The “Conditions” tab will list the conditions that must be met before your proposed measures may be considered and evaluated for suitability as endorsed voluntary consensus standards. You are asked to acknowledge reading and accepting the conditions.

**Can I make changes to a form once I have submitted it?** No. Once you submit your measure, you will NOT be able to return to this submission form to make further revisions. You will need to contact project staff.

**What if I need additional help?** Please contact the project staff at PQMsupport@battelle.org if you have questions regarding the information requested or submitting supplemental materials.

**NOTE: All measure submissions should be 508-compliant. Refer to the Checklist for Developer 508 Guidelines (PDF) to ensure all guidelines apply to all parts of your submission, including all fields and attachments used within the measure submission form.**

Please email us at PQMsupport@battelle.org if you experience technical difficulties using the online submission form.

Thank you for your interest in submitting measures to Battelle.

## Previous Submission Information (1 – 4)

**1) Select whether this measure was previously submitted to the prior consensus-based entity (the National Quality Forum [NQF]) and given an identifying number.**

☒ Previously submitted to NQF
☐ New measure, never submitted.

**2) Provide the measure number of the previously submitted measure.**

#3614, submitted for Spring 2021 Cycle
#3746, submitted for Spring 2023 Cycle

**3) If the measure has an electronic clinical quality measure (eCQM) version, provide the measure number of the previously submitted measure.**

Not applicable.

**4) If this eCQM has a registry version, provide the measure numbers of the previously submitted measure.**

Not applicable.

## Conditions (1 - 2)

**Several conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards. If any of the conditions are not met, the measure will not be accepted for consideration.**

A. A Measure Steward Agreement is signed or the steward is a government organization. (All non-government organizations must sign a Measure Steward Agreement.) For more information about completing a Measure Steward Agreement, please go to: Endorsement | Partnership for Quality Measurement (p4qm.org) and follow the instructions.

B. The measure owner/steward verifies there is an identified responsible entity and a process to maintain and update the measure on a schedule that is commensurate with the rate of clinical innovation, but at least every three years.

C. The intended use of the measure includes both accountability applications (including public reporting) and performance improvement to achieve high-quality, efficient healthcare.

D. The measure is fully specified and tested for reliability and validity.

E. The measure developer/steward attests that harmonization with related measures and issues with competing measures have been considered and addressed, as appropriate.

F. The requested measure submission information is complete and responsive to the questions so that all the information needed to evaluate all criteria is provided.

**1) Check if either of the following apply.**

☐ Proprietary measure or components (e.g., risk model, codes)
☐ Proprietary measure or components with fees
☒ None of the above

**2) Check the box below to agree to the conditions listed above.**

☒ I have read and accept the conditions as specified above

## Specifications: Maintenance Update (spma.01 - spma.02)

**spma.01) Indicate whether there are changes to the specifications since the last updates/submission. If yes, update the specifications in the Measure Specifications section of the Measure Submission Form, and explain your reasoning for the changes below.**

☒ No
☐ Yes

**spma.02) Briefly describe any important changes to the measure specifications since the last measure update and provide a rationale.**

**For annual updates, please explain how the change in specifications affects the measure results. If a material change in specification is identified, data from re-testing of the measure with the new specifications is required for early maintenance review.**

*For example, specifications may have been updated based on suggestions from a previous measure endorsement review.*

## Measure Specifications (sp.01 - sp.32)

**sp.01) Provide the measure title.**

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

**sp.02) Provide a brief description of the 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 incidence rate, reflecting days 0-30 days] minus expected [long-term incidence rate, reflecting days 91-360]).

**sp.03) Provide a rationale for why this measure must be reported with other measures to appropriately interpret results.**

Not applicable.

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

☐ Behavioral Health
☐ Behavioral Health: Alcohol, Substance Use/Abuse
☐ Behavioral Health: Anxiety
☐ Behavioral Health: Attention Deficit Hyperactivity Disorder (ADHD)
☐ Behavioral Health: Bipolar Disorder
☐ Behavioral Health: Depression
☐ Behavioral Health: Domestic Violence
☐ Behavioral Health: Other Serious Mental Illness
☐ Behavioral Health: Post-Traumatic Stress Disorder (PTSD)
☐ Behavioral Health: Schizophrenia
☐ Behavioral Health: Suicide
☐ Cancer
☐ Cancer: Bladder
☐ Cancer: Breast
☐ Cancer: Colorectal
☐ Cancer: Gynecologic
☐ Cancer: Hematologic
☐ Cancer: Liver
☐ Cancer: Lung, Esophageal
☐ Cancer: Prostate
☐ Cancer: Renal
☐ Cancer: Skin

☐ Cancer: Thyroid
☐ Cardiovascular
☐ Cardiovascular: Arrythmia
☐ Cardiovascular: Congestive Heart Failure
☐ Cardiovascular: Coronary Artery Disease
☐ Cardiovascular: Coronary Artery Disease (AMI)
☐ Cardiovascular: Coronary Artery Disease (PCI)
☐ Cardiovascular: Hyperlipidemia
☐ Cardiovascular: Hypertension
☐ Cardiovascular: Secondary Prevention
☐ Critical Care
☐ Critical Care: Assisted Ventilation
☐ Critical Care: Intensive Monitoring
☐ Dental
☐ Dental: Caries
☐ Dental: Tooth Loss
☒ Ears, Nose, Throat (ENT)
☐ Ears, Nose, Throat (ENT): Ear Infection
☐ Ears, Nose, Throat (ENT): Hearing
☐ Ears, Nose, Throat (ENT): Pharyngitis
☐ Ears, Nose, Throat (ENT): Tonsilitis
☐ Endocrine
☐ Endocrine: Calcium and Metabolic Bone Disorders
☐ Endocrine: Diabetes
☐ Endocrine: Female and Male Endocrine Disorders
☐ Endocrine: Hypothalamic-Pituitary Disorders
☐ Endocrine: Thyroid Disorders
☐ Eye Care
☐ Eye Care: Age-related macular degeneration (AMD)
☐ Eye Care: Cataracts
☐ Eye Care: Diabetic retinopathy
☐ Eye Care: Glaucoma
☐ Gastrointestinal (GI)
☐ Gastrointestinal (GI): Constipation
☐ Gastrointestinal (GI): Gall Bladder Disease
☐ Gastrointestinal (GI): Gastroenteritis
☐ Gastrointestinal (GI): Gastro-Esophageal Reflux Disease (GERD)
☐ Gastrointestinal (GI): Hemorrhoids
☐ Gastrointestinal (GI): Hernia
☐ Gastrointestinal (GI): Inflammatory Bowel Disease
☐ Gastrointestinal (GI): Irritable Bowel Syndrome
☐ Gastrointestinal (GI): Peptic Ulcer
☐ Genitourinary (GU)
☐ Genitourinary (GU): Benign Prostatic Hyperplasia
☐ Genitourinary (GU): Erectile Dysfunction/Premature Ejaculation
☐ Genitourinary (GU): Incontinence/pelvic floor disorders

☐ Genitourinary (GU): Prostatitis
☐ Genitourinary (GU): Urinary Tract Injection (UTI)
☐ Gynecology (GYN)
☐ Gynecology (GYN): Abnormal bleeding
☐ Gynecology (GYN): Endometriosis
☐ Gynecology (GYN): Infections
☐ Gynecology (GYN): Menopause
☐ Gynecology (GYN): Pelvic Pain
☐ Gynecology (GYN): Uterine fibroids
☐ Infectious Diseases (ID)
☐ Infectious Diseases (ID): HIV/AIDS
☐ Infectious Diseases (ID): Influenza
☐ Infectious Diseases (ID): Lyme Disease
☐ Infectious Diseases (ID): Meningococcal Disease
☐ Infectious Diseases (ID): Pneumonia and respiratory infections
☐ Infectious Diseases (ID): Sepsis
☐ Infectious Diseases (ID): Sexually Transmitted
☐ Infectious Diseases (ID): Tuberculosis
☐ Liver
☐ Liver: Viral Hepatitis
☐ Musculoskeletal
☐ Musculoskeletal: Falls and Traumatic Injury
☐ Musculoskeletal: Gout
☐ Musculoskeletal: Joint Surgery
☐ Musculoskeletal: Low Back Pain
☐ Musculoskeletal: Osteoarthritis
☐ Musculoskeletal: Osteoporosis
☐ Musculoskeletal: Rheumatoid Arthritis
☐ Neurology
☐ Neurology: Alzheimer's Disease
☐ Neurology: Autism
☐ Neurology: Brain Injury
☐ Neurology: Epilepsy
☐ Neurology: Migraine
☐ Neurology: Parkinson's Disease
☐ Neurology: Spinal Cord Injury
☒ Neurology: Stroke/Transient Ischemic Attack (TIA)
☒ Other (please specify here: Diagnostic Error)
☐ Palliative Care and End-of-Life Care
☐ Palliative Care and End-of-Life Care: Advanced Directives
☐ Palliative Care and End-of-Life Care: Amyotrophic Lateral Sclerosis (ALS)
☐ Palliative Care and End-of-Life Care: Hospice Management
☐ Palliative Care and End-of-Life Care: Inappropriate use of acute care services
☐ Palliative Care and End-of-Life Care: Pain Management
☐ Perinatal Health
☐ Perinatal Health: Labor and Delivery

☐ Perinatal Health: Newborn Care
☐ Perinatal Health: Post-Partum Care
☐ Perinatal Health: Preconception Care
☐ Perinatal Health: Prenatal Care
☐ Renal
☐ Renal: Acute Kidney Injury
☐ Renal: Chronic Kidney Disease (CKD)
☐ Renal: End Stage Renal Disease (ESRD)
☐ Renal: Infections
☐ Reproductive Health
☐ Reproductive Health: Family planning and contraception
☐ Reproductive Health: Infertility
☐ Reproductive Health: Male reproductive health
☐ Respiratory
☐ Respiratory: Acute Bronchitis
☐ Respiratory: Allergy
☐ Respiratory: Asthma
☐ Respiratory: Chronic Obstructive Pulmonary Disease (COPD)
☐ Respiratory: Dyspnea
☐ Respiratory: Pneumonia
☐ Respiratory: Sleep Apnea
☐ Surgery
☐ Surgery: Cardiac Surgery
☐ Surgery: Colorectal
☐ Surgery: Neurosurgery / Spinal
☐ Surgery: Orthopedic
☐ Surgery: Orthopedic Hip/Pelvic Fractures
☐ Surgery: Pediatric
☐ Surgery: Perioperative and Anesthesia
☐ Surgery: Plastic
☐ Surgery: Thoracic Surgery
☐ Surgery: Trauma
☐ Surgery: Vascular Surgery

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

☐ Access to Care
☐ Care Coordination
☐ Care Coordination: Readmissions
☐ Care Coordination: Transitions of Care
☐ Disparities Sensitive
☐ Health and Functional Status
☐ Health and Functional Status: Change
☐ Health and Functional Status: Nutrition
☐ Health and Functional Status: Obesity

☐ Health and Functional Status: Physical Activity
☐ Health and Functional Status: Quality of Life
☐ Health and Functional Status: Total Health
☐ Immunization
☒ Other (please specify here:  Safety: Diagnostic Error)
☐ Person-and Family-Centered Care: Person-and Family-Centered Care
☐ Person-and Family-Centered Care: Workforce
☐ Primary Prevention
☐ Primary Prevention: Nutrition
☐ Primary Prevention: Tobacco Use
☒ Safety
☐ Safety: Complications
☐ Safety: Healthcare Associated Infections
☐ Safety: Medication
☐ Safety: Overuse
☐ Screening

**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.*

☒ Adults (Age >= 18)
☐ Children (Age < 18)
☐ Elderly (Age >= 65)
☐ Populations at Risk: Dual eligible beneficiaries of Medicare and Medicaid
☐ Populations at Risk: Individuals with multiple chronic conditions
☐ Populations at Risk: Veterans
☐ Women

**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.*

☐ Accountable Care Organization
☐ Clinician: Group/Practice
☐ Clinician: Individual
☒ Facility
☐ Health Plan
☐ Integrated Delivery System
☐ Other (please specify here:  )
☐ Population: Community, County or City
☐ Population: Regional and State

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

Check ONLY the settings for which the measure is SPECIFIED and TESTED.
☒ Ambulatory Care (Emergency Department)

☐ Behavioral Health
☐ Home Care
☐ Inpatient/Hospital
☐ Other (please specify here: )
☐ Outpatient Services
☐ Post-Acute Care

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

https://www.hopkinsmedicine.org/armstrong_institute/centers/center_for_diagnostic_excellence/dizzy-stroke-ed-specs.html?L

**sp.10) Indicate whether Health Quality Measure Format (HQMF) specifications are attached.**

*Attach the zipped output from the measure authoring tool (MAT) for eCQMs - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications). HQMF specifications are attached.*

☒ HQMF specifications are NOT attached (Please explain).

Not an eCQM measure

**sp.11) Attach the simulated testing attachment.**

*All eCQMs require a simulated testing attachment to confirm that the HTML output from Bonnie testing (or testing of some other simulated data set) includes 100% coverage of measured patient population testing, with pass/fail test cases for each sub-population. This can be submitted in the form of a screenshot.*

☐ Testing is attached
☒ Testing is NOT attached (please explain)

Not an eCQM measure

**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 at PQMsupport@battelle.org. Provide descriptors for any codes. Use one file with multiple worksheets, if needed.*

☒ Available in attached Excel or csv file
☐ No data dictionary/code table – all information provided in the submission form

For the question below: state the outcome/process being measured. Calculations of the risk-adjusted outcome measures should be described in sp.22.

**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.*

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 hospital discharge diagnosis of stroke.

For the question below: describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in sp.22.

**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.*

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.

For the question below: state the target population for the outcome. Calculation of the risk-adjusted outcome should be described in sp.22.

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

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

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."

For the question below: describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in sp.22.

**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.*

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.

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

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

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.

**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.*

Not applicable.

**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.*

Not applicable.

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

☐ Yes
☒ No

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

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

☒ No risk adjustment or risk stratification
☐ Statistical risk model
☐ Stratification by risk category/subgroup (specify number of risk factors)
☐ Other approach to address risk factors (please specify here: )

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

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

☐ Categorical, e.g., yes/no
☐ Continuous variable, e.g. average
☐ Count
☐ Frequency Distribution
☐ Non-weighted score/composite/scale
☐ Other (please specify here: )
☒ Rate/proportion
☐ Ratio
☐ Weighted score/composite scale

**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.***

☐ Better quality = Higher score
☒ Better quality = Lower score
☐ Better quality = Score within a defined interval
☐ Passing score defines better quality

**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.*

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 hospital 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 hospital 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.

**sp.25) Attach a copy of the instrument (e.g. survey, tool, questionnaire, scale) used as a data source for your measure, if available.**

☐ Copy of instrument is attached.
☒ Copy of instrument is NOT attached (please explain).

Not applicable.

**sp.26) Indicate the responder for your instrument.**

☐ Patient
☐ Family or other caregiver
☐ Clinician
☐ Other (specify)

Not applicable.

**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 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.*

Not applicable.

**sp.28) Identify whether and how proxy responses are allowed.**

Not applicable.

**sp.29) Survey/Patient-reported data.**

*Provide instructions for data collection and guidance on minimum response rate. Specify calculation of response rates to be reported with performance measure results.*

Not applicable.

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

☐ Assessment Data
☒ Claims
☐ Electronic Health Data
☐ Electronic Health Records
☐ Instrument-Based Data
☐ Management Data
☐ Other (please specify here: )
☐ Paper Medical Records
☐ Registry Data

**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.*

Not applicable.

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

☐ Available at measure-specific web page URL identified in sp.09
☐ Available in attached appendix in Question 1 of the Additional Section
☒ No data collection instrument provided

## Importance to Measure and Report: Maintenance of Endorsement (1ma.01)

**1ma.01) Indicate whether there is new evidence about the measure since the most recent maintenance evaluation. If yes, please briefly summarize the new evidence, and ensure you have updated entries in the Evidence section as needed.**

☐ Yes
☐ No

Initial submission of the measure.

## Importance to Measure and Report: Evidence (Complete for Outcome Measures) (1a.01 - 1a.03)

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:**

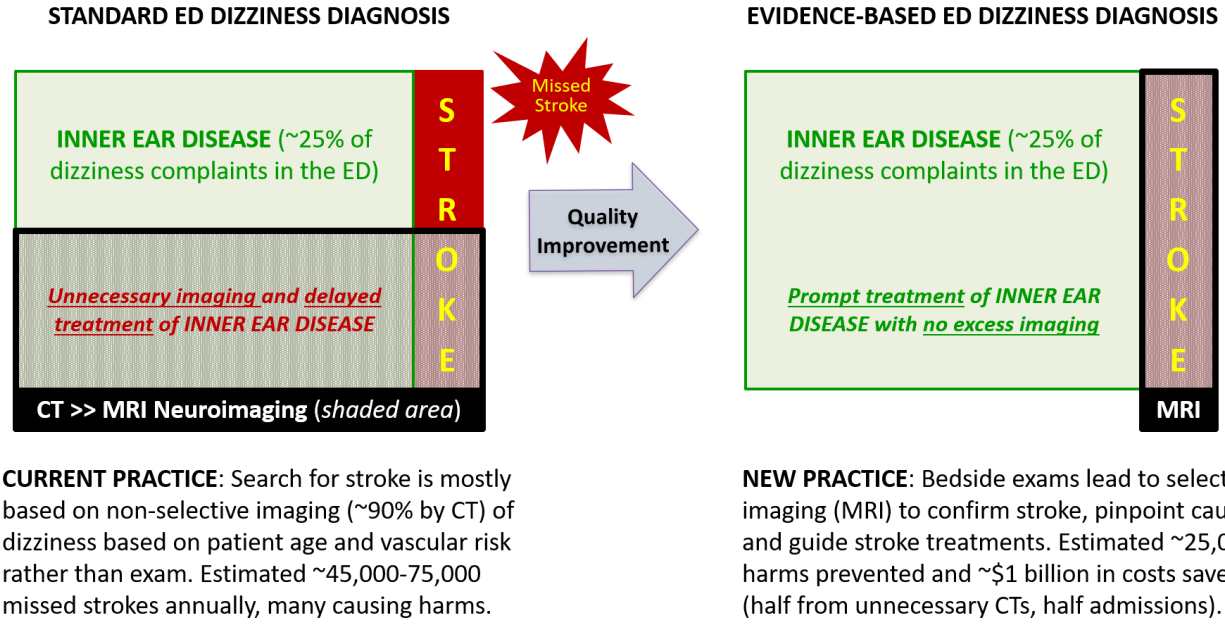
**1a.01) Provide a logic model.**

*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.*

> ***This measure is conceptualized using the well-established "SPADE" method for measuring diagnostic errors. For more details on the SPADE method, please see the APPENDIX.***

***A. Below is a description of the stepwise mechanism by which this proposed measure will improve quality/safety for patients…***

1. This measure reflects missed strokes in ED patients presenting with dizziness or vertigo
2. Improvement on the measure requires QI efforts that improve diagnosis of ED patients that present with dizziness/vertigo (Figure 1)
3. These QI efforts will improve diagnosis both for patients with stroke and patients with inner ear disease
4. Benefits to patients will accrue from the prompt application of research-proven treatments
    a. Those with stroke will benefit from tPA or early secondary prevention, as appropriate
    b. Those with benign paroxysmal positional vertigo will benefit from prompt canalith repositioning and less CT radiation
5. These benefits to stroke patients (4a), in turn, will result in a "better" measure score (Figure 2)




**CURRENT PRACTICE**: Search for stroke is mostly based on non-selective imaging (~90% by CT) of dizziness based on patient age and vascular risk rather than exam. Estimated ~45,000-75,000 missed strokes annually, many causing harms.

**NEW PRACTICE**: Bedside exams lead to selective imaging (MRI) to confirm stroke, pinpoint cause, and guide stroke treatments. Estimated ~25,000 harms prevented and ~$1 billion in costs saved (half from unnecessary CTs, half admissions).

Figure 1. Theory for ED practice change. Standard practice in diagnosing dizziness now rests largely on CT to search for stroke in older patients with vascular risk factors. However, CT is ineffective for diagnosing vestibular strokes. Because inner ear causes are also more common among older populations with stroke risk factors, imaging is overused in inner ear diseases. Simultaneously, young patients (or old patients without vascular risk factors) who *do* have strokes as the cause may inadvertently be sent home untreated, sometimes with devastating consequences.[1,2] QI interventions such as teleconsultation will focus neuroimaging on directing stroke treatments, and more patients with inner ear disease will be correctly diagnosed and treated, preventing unnecessary imaging and admission.
*Abbreviations: CT – computerized tomography; MRI – magnetic resonance imaging; QI – quality improvement*

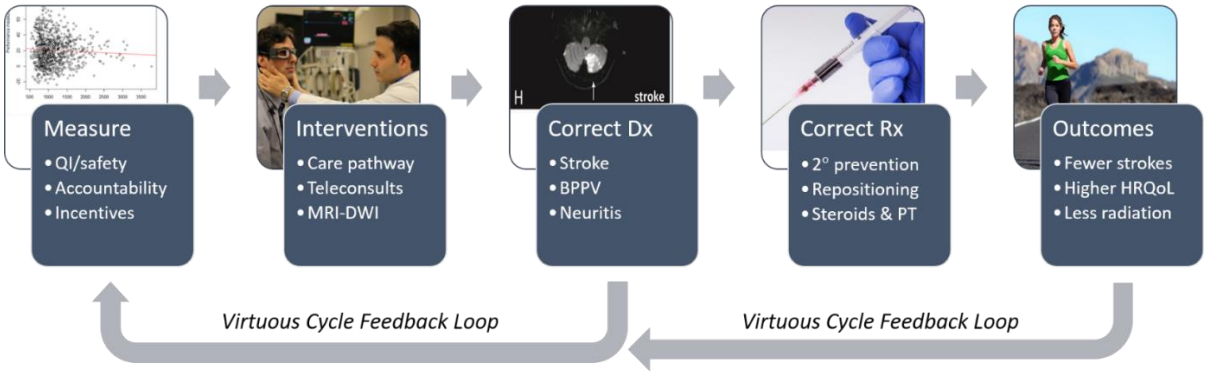



Figure 2. Logic model by which proposed measure will improve quality and safety for patients.

*Abbreviations: 2° – secondary; Dx – diagnosis; HRQoL – health-related quality of life; MRI-DWI – magnetic resonance imaging with diffusion weighted images; PT – physical therapy; QI – quality improvement; Rx – treatment*

### *B. Evidence supporting a positive impact of the measure on patient care…*

1. **SYSTEMATIC REVIEW EVIDENCE THAT BETTER EYE EXAMS INCREASE CLINICAL DIAGNOSTIC ACCURACY**: There is strong evidence from multiple systematic reviews with meta-analyses of multiple prospective observational studies that bedside eye movement exams ("HINTS") in the hands of neurologists can more accurately diagnose stroke in patients with dizziness than even MRI scans.[3–6] Furthermore, the accuracy of these bedside exams far exceeds that of the more commonly used imaging technique of CT (which misses over 90% of acute posterior fossa strokes presenting with dizziness [reviewed in Newman-Toker, 2016[7]]), as well as the overall accuracy of current ED care, in which 40% of strokes presenting with dizziness are missed.[8] Neurology consultation services directly to the ED have demonstrated dramatically improved diagnostic accuracy, while simultaneously reducing inappropriate imaging.[9,10] Reductions in inappropriate CT use eliminate unnecessary irradiation, thereby cutting cancer risk, so improving outcomes for patients.[11] And while untrained ED clinicians do not perform this bedside testing well, those who are trained using direct observation and feedback methods achieve similar diagnostic results to those obtained by specialists—(sensitivity: 92.9% [95% CI 70–100%]; specificity: 96.4% [95% CI 93–98%]; positive predictive value: 81.3% [95% CI 61–87%]; negative predictive value: 98.8% [95% CI 95–100%]).[12,13] Furthermore, a recently published guideline from the Society for Academic Emergency Medicine, developed using rigorous GRADE guideline methods, endorses and supports the use of these diagnostic methods in the ED.[14]

2. **FACE VALIDITY THAT BETTER DIAGNOSIS YIELDS BETTER TREATMENT**: It is face valid that increasing correct diagnosis of posterior stroke in patients with dizziness and vertigo will lead to greater application of randomized trial and guideline approved stroke treatments in the ED. Likewise the same for inner ear diseases.

3. **RCT EVIDENCE THAT EARLY TREATMENT OF MINOR STROKE/TIA IMPROVES OUTCOMES**: It is proven through randomized clinical trials (CHANCE, POINT) that certain patients with TIA and minor stroke benefit from the application of early secondary prevention treatments, such as dual antiplatelet therapy. Combined results in over 10,000 patients show that treatment in the first 24 hours cuts the risk of a major stroke by 34% in the next 21 days.[15] Other empirical studies of the benefit of immediate stroke treatments include the following—preventable adverse outcomes of misdiagnosis result from missed opportunities for thrombolysis,[16,17] early surgery for malignant posterior fossa edema,[18,19] or prevention of subsequent infarction.[20–22] Rapid treatment improves health outcomes[23,24] and prompt prophylaxis lowers repeat stroke risk by up to 80%.[25,26] Thus, patients generally benefit from early, correct diagnosis.

4. **RCT EVIDENCE THAT EARLY TREATMENT OF INNER EAR DISEASES IMPROVES OUTCOMES**: Benefits also accrue to patients with dizziness or vertigo who are correctly diagnosed with inner ear disease (benign paroxysmal positional vertigo and vestibular neuritis) who receive guideline-supported treatments with randomized controlled trial evidence,[27–33] and direct harms of misdiagnosis[34] are reduced.

5. **FACE VALIDITY THAT PREVENTING MAJOR STROKES WILL LOWER THE MEASURE SCORE**: It is face valid that if there are fewer subsequent major strokes among those treated, then there will be fewer short-term hospitalizations for stroke, which is, in turn, reflected in the measure (i.e., by reducing the “n” in the numerator). Furthermore, properly identifying such patients in the first place will remove these higher risk patients from the denominator (by correctly diagnosing stroke rather than “benign” inner ear disease or non-specific dizziness); this will tend to lower the observed number of subsequent strokes towards the expected population base rate (which is included as part of the measure calculation, which is observed minus expected).

***C. Evidence of improved diagnostic accuracy in clinical practice with consult-based quality improvement…***

Recent data (Table 1) from a quality improvement intervention (Tele-Dizzy) involving remote neurology consultations show dramatic *increases* in both stroke and specific inner ear diagnoses, along with dramatic *decreases* in inappropriate imaging among 287 patients who underwent consultation, relative to a matched emergency department population. These results provide compelling ***empirical evidence*** supporting the link between a ***healthcare intervention/service*** and the outcome of improved diagnosis, as well as better patient outcomes (reduction in unnecessary irradiation). It is inferentially logical and face valid, then, that these results, implemented more broadly, could be measured using this measure.

**Table 1. Results of Tele-Dizzy Quality Improvement Intervention at Johns Hopkins Hospital (n=287 tele-consults).**

| Category | Parameter | Baseline* | Tele-Dizzy | Improvement | p-value ($\chi^2$) |
|---|---|---|---|---|---|
| Diagnostic Yield | Specific Vestibular Diagnosis Rate | 77 (20.6%) | 163 (56.8%) | ↑ 176% | P<0.0001 |
| | Stroke Diagnosis Rate | 1 (0.3%) | 20 (7.0%) | ↑ 2,506% | P<0.0001 |
| | Non-Diagnosis Rate | 235 (62.8%) | 86 (30.0%) | ↓ 52% | P<0.0001 |
| Test Utilization | Neuroimaging (CT or MRI) | 198 (52.9%) | 70 (24.4%) | ↓ 54%† | P<0.0001 |
| Patient Outcomes | Excess 30-day stroke hospitalizations | 0.1%‡ | 0 (0.0%)‡ | ↓ 100%‡ | NA |

** Baseline rates for diagnostic accuracy and test utilization are from 374 ED patients with a presenting symptom of dizziness (seen outside of Tele-Dizzy consultation hours) who had mention of "nystagmus" in notes and were comparable on the variables age, sex, and ED triage acuity.*

*† CT scans were reduced by 96% (from 49.2% to 1.7%, p<0.0001) and MRIs for patients without strokes were unchanged (15.5% vs. 15.7%, p=0.95).*

*‡ Baseline 30d stroke hospitalizations are calculated as in Measure #3746 (not using the comparator population for Tele-Dizzy, which was too small for a precise estimate). The Tele-Dizzy value is based on actual patients seen at the same hospital – thus far, there have been zero stroke returns.*

### *D. Summary*

The logic model above offers a set of valid logical links between the measure, quality improvement interventions, and improved patient outcomes. Each of the key steps is either supported by strong empirical evidence or has face validity.

**1a.02) Provide evidence that the target population values the measured outcome, process, or structure and finds it meaningful.**

*Describe how and from whom input was obtained.*

***Patients and ED Clinicians are Concerned about this Issue***

For patients seeking care in the ED with new dizziness, worry about the cause can be prominent, often including fear of having a stroke. After dangerous causes are excluded, patient focus often shifts quickly to treatment. The ED physician's main diagnostic focus is aligned with the patient's concern—in dizziness or vertigo without clear medical or neurological features, their goal is to distinguish "central" (brain) from "peripheral" (inner ear) causes.[35] In a 2005 survey of over 1,000 ED physicians, "identification of central or serious vertigo" was the #1 desired decision rule for adults.[36] In 2013, 95% of over 350 ED physicians wanted a valid approach to "help decide whether to obtain neuroimaging" or "exclude stroke as a cause of dizziness in ED patients without neuroimaging."[37] Even after stroke is excluded as a cause, patients often do not receive a specific/correct diagnosis, optimal treatment, or appropriate referral for their undiagnosed (or misdiagnosed) inner ear disease.[38] Patients attest to the negative consequences to their quality of life (Figure 3).

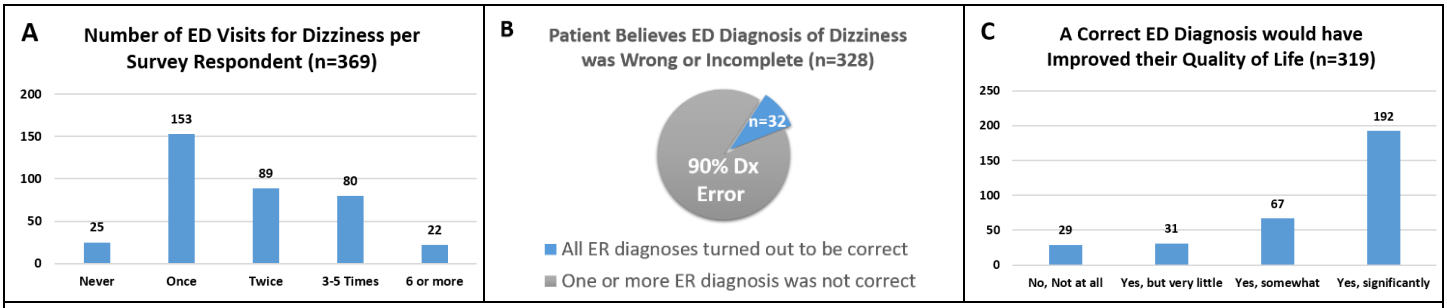



Figure 3. Internet survey of 373 patients with vestibular disorders, conducted by stakeholder partner the Vestibular Disorders Association (VEDA) in preparation for an upcoming trial. Results clearly show that for these patients, most visit the ED (93%), the majority more than once (*Panel A*); patient-reported diagnostic errors are the norm (90%) (*Panel B, Dx=diagnostic*); and patients believe that more accurate initial ED diagnoses would have improved their quality of life (91%) (most believed the quality-of-life difference would have been significant) (*Panel C*).

***ED Clinicians are Supportive of this Measurement Approach***

Our AHRQ-funded project, "Towards a National Diagnostic Excellence Dashboard—Partnering with Stakeholders to Construct Evidence-Based Operational Measures of Misdiagnosis-Related Harms" (PI Newman-Toker, R01 HS 27614, 2020-2024) is explicitly **designed to engage stakeholder partners, including patients and emergency physicians, in measure co-creation to optimize the final measure specifications.**

Regular meetings have been conducted over the past three years with partner institutions the Society to Improve Diagnosis in Medicine (SIDM) and the American College of Emergency

Physicians (ACEP). A technical expert panel (TEP) includes representation from frontline ED clinicians (physicians, nurses, advanced practice providers), emergency medical services, hospitalists, vestibular neurologists, compliance experts, risk management, hospital quality and safety, hospital administration, measurement experts, data science and analytics experts, as well as payers, purchasers, and experts in policymaking. The TEP has held five meetings between June 2021 and November 2022. Each of the nine attributes slated for quality measure refinement was presented with supporting data and expected tradeoffs for TEP discussion. Further TEP discussions included survey development and deployment, survey results review, and recommendations for subsequent survey work, in addition to a session focused on considering solutions.

A draft stakeholder survey was developed by the research team for distribution to the ACEP E-QUAL network (targeting both ED clinicians and ED medical directors or quality/safety officers). The research team, in stakeholder partnership with SIDM and ACEP, distributed the survey in early spring 2022.

The survey was completed by 31 ED front-line clinicians and by 36 ED directors. For both front-line clinicians and directors, a strong majority (67%) worked in an academic ED and over 85% said they worked at a designated stroke center (whether primary or comprehensive).

Both groups (frontline clinicians 81%, medical directors 85%) said that receiving hospital/ED-level feedback on missed stroke in dizziness/vertigo presentations would improve their practice and the quality of care for patients with dizziness/vertigo, and over 90% of both groups said they would welcome such feedback.

**1a.03) Provide empirical data demonstrating the relationship between the outcome (or PRO) and at least one healthcare structure, process, intervention, or service.**

**Relationship between the Outcome (*Stroke Misdiagnosis*) and a Healthcare Process (*Appropriate Neuroimaging of Dizzy Patients*)**

We used the full national Medicare fee-for-service dataset to explore the relationship between a hospital's performance on the dizzy-stroke measure and its use of imaging in dizzy patients and the type of imaging used (CT vs. MRI). We performed two types of analyses: (1) facility-level analyses; (2) visit-level analyses. For facility-level analyses, we first classified facilities into high-imaging-rate facilities and low-imaging-rate facilities and then calculated performance on the measure for the two groups of patients treated at these two types of facilities. A cutoff for defining what a high versus low imaging rate was chosen such that the number of patients in the two groups were approximately the same. For visit-level analyses, we grouped individual patients based on whether they were imaged at index visits and calculated the measure for each group.

To strengthen our understanding of measure precision/reliability when a more comprehensive dataset is used for its calculation, we used the HCUP SID and SEDD datasets for Florida (2016-2019) to calculate Florida hospitals' performance on the measure. The SID and SEDD datasets include claims for all ED and hospital discharges, which expands our sample size by a factor of four (Medicare fee-for-service represents about 25% of patients in an average U.S. hospital) without increasing the duration of the metric's performance window (a 3-year rolling window).

For the imaging analysis, a general conclusion from the facility-level analysis (Table 2) is that a higher imaging rate of any type is associated with a lower rate of misdiagnosis, particularly if the imaging is by MRI. A general conclusion from the visit-level analysis (Table 3) is that misdiagnosis increases with the use of CT and decreases with the use of MRI. Taken together, ***these two findings suggest that appropriate use of MRIs (whether alone or in combination with CT scans) appear to represent a potent intervention that hospitals might use to improve their diagnostic performance of stroke patients who present with dizziness or vertigo symptoms.***

**Table 2. Facility-level comparison of stroke quality metric by facility imaging frequency.**

| Imaging Type | Facilities with LOWER Rates of Imaging | | | | Facilities with HIGHER Rates of Imaging | | | | *P*-value * |
|---|---|---|---|---|---|---|---|---|---|
| | Imaging Range (%) | Facilities (N) | Index Visits (N) | Stroke Metric (O-E)† | Imaging Range (%) | Facilities (N) | Index Visits (N) | Stroke Metric (O-E)† | |
| Any CT or MRI | 0-56 | 3463 | 1,039,539 | **27.0** (25.8, 28.3) | 56-100 | 1910 | 1,043,517 | **23.3** (22.1, 24.5) | **< 0.01** |
| Any CT | 0-51 | 3339 | 1,038,760 | **26.3** (25.0, 27.5) | 51-100 | 2034 | 1,044,296 | **24.0** (22.8, 25.3) | **< 0.01** |
| Only CT no MRI | 0-51 | 3339 | 1,039,020 | **26.3** (25.0, 27.5) | 51-100 | 2034 | 1,044,036 | **24.1** (22.9, 25.3) | **< 0.01** |
| Any MRI | 0-3 | 3769 | 1,043,988 | **29.2** (27.9, 30.5) | 3-100 | 1604 | 1,039,068 | **21.1** (19.9, 22.2) | **< 0.01** |
| Only MRI no CT | 0-3 | 3768 | 1,044,097 | **29.2** (27.9, 30.5) | 3-100 | 1605 | 1,038,959 | **21.1** (20.0, 22.3) | **< 0.01** |

* *P*-value for comparison between the stroke misdiagnosis metric at LOWER vs. HIGHER-imaging facilities. Note that the greater differences in the stroke misdiagnosis metric occur when MRI is involved rather than when CT is involved.

† The stroke metric is based on an "O-E" (observed minus expected) rate of primary stroke hospitalizations within 30 days after an ED treat-and-release visit with a non-specific or benign dizziness/vertigo diagnosis code per 10,000 visits. The benefit to adverse stroke outcomes is ~2 per 10,000 visits for CT vs. no imaging and ~8 per 10,000 visits for MRI vs. no imaging. The difference between CT and MRI (~6 per 10,000), without any other improvements in diagnosis, would translate to a difference of ~3,000 preventable harms (stroke hospitalizations) each year in the US. As noted below in the footnotes to Table 3, however, this could easily be a 2.5-fold underestimate of the impact.

**Table 3. Visit-level comparison of stroke quality metric by presence of imaging at ED index visit.**

| Imaging Type | No Imaging # Visits (N) | No Imaging Stroke Metric (O-E)* | Yes Imaging # Visits (N) | Yes Imaging Stroke Metric (O-E)* | P-value† |
|---|---|---|---|---|---|
| Any CT or MRI | 1,032,738 | 23.2 (22.0, 24.4) | 1,048,135 | 27.1 (25.8, 28.4) | < 0.01 |
| Any CT | 1,126,734 | 22.6 (21.5, 23.7) | 954,139 | 28.2 (26.9, 29.5) | < 0.01 |
| Only CT, no MRI | 1,128,123 | 22.6 (21.5, 23.8) | 952,750 | 28.1 (26.8, 29.5) | < 0.01 |
| Any MRI | 1,985,488 | 25.6 (24.7, 26.5) | 95,385 | 16.7 (13.0, 20.4) | < 0.01 |
| Only MRI, no CT | 1,986,877 | 25.6 (24.7, 26.5) | 93,996 | 15.8 (12.1, 19.4) | < 0.01 |

* The stroke metric is based on an "O-E" (observed minus expected) rate of primary stroke hospitalizations within 30 days after an ED treat-and-release visit with a non-specific or benign dizziness/vertigo diagnosis code per 10,000 visits. The difference in adverse stroke outcomes is ~5.5 ***worse*** per 10,000 visits for CT vs. no imaging and ~9.8 *better* per 10,000 visits for MRI vs. no imaging. This comports with prior data showing that negative CTs in those with dizziness actually predict a higher risk of missed stroke,[39] presumably because of correct selection of higher risk patients, followed by false reassurance by a "normal" CT scan, which has very low sensitivity for stroke. This argues that the positive impact of MRI is quite substantial, since the same selection effect present among patients undergoing CT tends to bias towards the null the difference between the "no imaging" and "yes MRI" visits, yet we still see a large beneficial effect of MRI. At a minimum, this indicates a difference of more than 15 per 10,000 visits with MRI, which could translate to ~7,500 preventable harms (stroke hospitalizations) each year in the US. This is without making any other improvements in diagnosis of patients with dizziness or vertigo. Since MRI misses about 25% of strokes presenting with dizziness or vertigo, the number of potentially preventable harms may be even higher.

† *P*-value for comparison between the stroke misdiagnosis metric for visits without vs. with imaging. Use of CT actually worsens the risk of stroke misdiagnosis as assessed by the quality metric, while MRI improves it.

Summary of Neuroimaging as a Process that Impacts the Outcome Measure:

- Imaging by MRI prevents adverse outcomes from missed strokes in dizziness/vertigo. This is despite the fact that MRI misses 25% of acute strokes causing dizziness.
- Imaging by CT is linked to an increased risk of adverse outcomes from missed stroke; this effect likely represents correct clinical risk stratification and false reassurance by falsely negative CT neuroimaging for acute ischemic stroke, common in dizziness. This is unsurprising given that CT sensitivity for posterior fossa ischemic strokes presenting with dizziness or vertigo may be as low as 7-16%.[7]
- This suggests that one possible intervention to reduce missed strokes is to obtain more MRIs and fewer CTs in appropriate patients with dizziness or vertigo**. It also fairly convincingly demonstrates that the quality measure would be sensitive to such an intervention (i.e., there is an existing way to "move the needle" on performance).**

## Importance to Measure and Report: Evidence (Complete for Process Measures) (1a.03 - 1a.16)

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.*

**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.*

☐ Clinical Practice Guideline recommendation (with evidence review)
☐ US Preventive Services Task Force Recommendation
☐ Other systematic review and grading of the body of evidence (e.g., Cochrane Collaboration, AHRQ Evidence Practice Center)
☐ Other (please specify here: )

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, you may add additional tables to the relevant sections. Please follow the 508 Checklist for tables.

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

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

**systematic review.**

**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.**

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

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

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

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

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

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

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

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

**Evidence**

**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.**

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

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

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

## Importance to Measure and Report: Gap in Care/Disparities (1b.01 - 1b.05)

**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.*

Diagnostic error is a major public health problem.[40] The lack of operational measures is a critical barrier to improving diagnostic quality.[41,42] Three major disease categories (vascular events, infections, and cancer) account for three-fourths of all serious harms from diagnostic error as identified by malpractice claims.[43] Among vascular events, missed stroke is the leading cause of serious harm to patients. Misdiagnosis of stroke disproportionately occurs when patients present with symptoms/signs that are not typical or obvious for stroke.[8,44] For example, the most common clinical presentation of missed stroke occurs when patients present with dizziness or vertigo, which can easily be mistaken for inner ear disease.[8] Annually in US emergency departments (ED), an estimated 45,000-75,000 patients that present with dizziness or vertigo and have strokes, are misdiagnosed and erroneously discharged from the ED.[44]

ED patients with acute dizziness and vertigo could be correctly diagnosed with stroke using evidence-based bedside examinations,[3,35] but there is a large evidence-practice gap[38] in ED diagnosis, resulting in substantial harms to patients.[44] Without a timely and accurate diagnosis, these patients suffer misdiagnosis-related harms[45] because they do not receive prompt treatment for this time-sensitive condition.[8] The most common harm is a preventable major stroke leading to a subsequent hospitalization after the patient has had a minor stroke or transient ischemic attack (TIA).[21,22] Crude short-term stroke hospitalization rates per 10,000 ED dizziness discharges vary at least from 20-80.[44] Adjusting for baseline stroke risk across groups does not eliminate practice variation.[46]

This outcome measure tracks the rate of missed strokes in the ED—i.e., patients admitted to the hospital for a stroke within 30 days of an ED discharge with a non-specific diagnosis of benign dizziness diagnosis or a specific inner ear/vestibular diagnosis (collectively referred to as “benign dizziness”). This measure is the first operationally viable performance measure of stroke misdiagnosis in the hospital setting. Hospital EDs will be able to use the measure to internally track their performance over time as they work to implement interventions to reduce stroke misdiagnosis. The measure can also be used by external entities for public reporting and pay-for-performance, as external pressure to encourage improvement in diagnostic quality.

**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.*

Current (1/1/2015-12/31/2017); Data Source: Medicare Fee-for-Service + Medicare Advantage; Number of Measured Entities: 967 Hospital EDs; Number of Patients: 383,017; Mean Score: 17.70; SD: 30.04; Min Score: (-29.15); Max Score: 165.32; IQ Range: (-7.32, 31.43); Median scores by decile: (-17.58, -12.10, -7.35, 0.00, 10.41, 16.91, 23.54, 31.44, 49.62, 73.66)

Past (1/1/2012-12/31/2014); Data Source: Medicare Fee-for-Service + Medicare Advantage; Number of Measured Entities: 965 Hospital EDs; Number of Patients: 371,788; Mean Score: 20.05; SD: 33.03; Min Score: (-38.02); Max Score: 162.90; IQ Range: (-7.97, 39.84); Median scores by decile: (-20.51, -13.12, -7.97, 2.41, 12.36, 19.04, 27.48, 39.84, 55.18, 76.68)

Past (1/1/2009-12/31/2011); Data Source: Medicare Fee-for-Service + Medicare Advantage; Number of Measured Entities: 804 Hospital EDs; Number of Patients: 295,678; Mean Score: 26.56; SD: 36.83; Min Score: (-41.93); Max Score: 219.94; IQ Range: (-0.10; 47.30); Median scores by decile: (-22.02, -13.04, -0.10, 9.28, 17.50, 24.61, 35.66, 47.30, 63.39, 93.58)

**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.**

Not applicable.

**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.*

Not available.

**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.**

Prior research has identified that women and minorities are at ~20-30% increased odds of stroke misdiagnosis and patients 18-44 years old are at roughly 7-fold increased odds.[47,48]

Differences by Gender

• Newman-Toker et al., Diagnosis, 2014[47]: Found the odds of a probable misdiagnosis were lower among men (OR 0.75) than women.
• Von Kleist et al., Neurology, 2019[49]: Found within misdiagnosed stroke/TIA patients (n=117), there was a significant difference between gender in initial diagnosis (p=0.0052). Females were more likely than males to be given an “uncertain” diagnosis (44.07% vs 17.24%).

Differences by Race
• Newman-Toker et al., Diagnosis, 2014[47]: Found the odds of a probable misdiagnosis were higher among Blacks (OR 1.18), Asian/Pacific Islanders (OR 1.29), and Hispanics (OR 1.30) than non-Hispanic Whites.

Differences by Age
• Kuruvilla et al, Journal of Stroke and Cerebrovascular Diseases, 2011[16]: Found patients age ≤35 years (P=.05) were more likely to be misdiagnosed.
• Newman-Toker et al., Diagnosis, 2014[47]: Found the odds of a probable misdiagnosis were lower among older individuals (using 18-44 years as the base); 45-64 years old (OR 0.43); 65-74 years old (OR 0.28); ≥75 years old (OR 0.19).

## Scientific Acceptability: Maintenance (2ma.01 - 2ma.04)

**2ma.01) Indicate whether additional empirical reliability testing at the accountable entity level has been conducted. If yes, please provide results in the following section, Scientific Acceptability: Reliability - Testing. Include information on all testing conducted (prior testing as well as any new testing).**

***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 Submission:**

Testing from the previous submission here.

☐ Yes
☐ No

**2ma.02) Indicate whether additional empirical validity testing at the accountable entity level has been conducted. If yes, please provide results in the following section, Scientific Acceptability: Validity - Testing. Include information on all testing conducted (prior testing as well as any new testing).**

***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 Submission:**

Testing from the previous submission here.

☐ Yes
☐ No

**2ma.03) For outcome, patient-reported outcome, resource use, cost, and some process measures, risk adjustment/stratification may be conducted. Did you perform a risk adjustment or stratification analysis?**

☐ Yes

☐ No

**2ma.04) For maintenance measures in which risk adjustment/stratification has been performed, indicate whether additional risk adjustment testing has been conducted since the most recent maintenance evaluation. This may include updates to the risk adjustment analysis with additional clinical, demographic, and social risk factors.**

**Please update the Scientific Acceptability: Validity - Other Threats to Validity section.**

**Note: This section must be updated even if social risk factors are not included in the risk adjustment strategy.**

☐ Yes - Additional risk adjustment analysis is included
☐ No additional risk adjustment analysis included

## Scientific Acceptability: Reliability - Testing (2a.01 - 2a.12)

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 Battelle staff at PQMsupport@battelle.org 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 Battelle staff at PQMsupport@battelle.org with any questions.
- 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 the 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

OR

• rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful 16 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 measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

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.**

☐ Assessment Data
☒ Claims
☐ Electronic Health Data
☐ Electronic Health Records
☐ Instrument-Based Data
☐ Management Data
☐ Other (please specify here: )
☐ Paper Medical Records
☐ Registry Data

**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).*

**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).

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

*Use the following format: "MM-DD-YYYY - MM-DD-YYYY"*

01-01-2015 - 12-31-2019

**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.*

☐ Accountable Care Organization
☐ Clinician: Group/Practice
☐ Clinician: Individual
☒ Facility
☐ Health Plan
☐ Integrated Delivery System
☐ Other (specify)
☐ Population: Community, County or City
☐ Population: Regional and State

**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.*

**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)[50] 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.[51] that looked at ED visit volume distribution across U.S. hospitals (with expected growth in visits during the last 10 years).

**Table 4. Comparison of ED Visits between 2010 study and our National Hospital Dataset.**

| ED Visits per year | Muelleman et al.[51] (2007 data)<br><br>N=4,874 Non-Federal EDs | Our Dataset (2017 data)<br><br>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% |
| 40,000-49,000 | 8% | 8% |
| >50,000 | 12% | 23% |

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[52]).

The aggregate distribution of visits in Florida EDs skewed a bit higher than the 2010 study by Muelleman et al.[51] 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.

**Table 5. Comparison of ED Visits between 2010 study and our State Hospital Dataset.**

| ED Visits per year | Muelleman et al.[51] (2007 data)<br><br>N=4,874 Non-Federal EDs | Our Dataset (2016-2019 data)<br><br>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% |
| >50,000 | 12% | 18% |

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.

**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.*

**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).

**Table 6. Percentage of Patients in National Hospital Dataset with Demographic Characteristic.**

| 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% |

**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,[53] as expected for Florida, which has a higher percentage of residents over age 65 than any state other than Maine.[54] The female-to-male distribution is typical for dizziness across age groups (roughly 60% female, 40% male).

**Table 7. Percentage of Patients in State Hospital Dataset with Demographic Characteristic.**

| 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% |
| • 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% |

**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.**

**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 July1, 2016 – June 30, 2017 were used for data-element validity testing.

**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.*

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,[47] 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.

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.
☐ Patient or Encounter-Level (e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements)

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

**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.*

Performance measure score reliability was calculated using signal-to-noise analysis as described in the technical report "The Reliability of Provider Profiling: A Tutorial," [55] by J.L. Adams, 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.

**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, Measure Evaluation Criteria).*

### *National hospital-level testing*

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

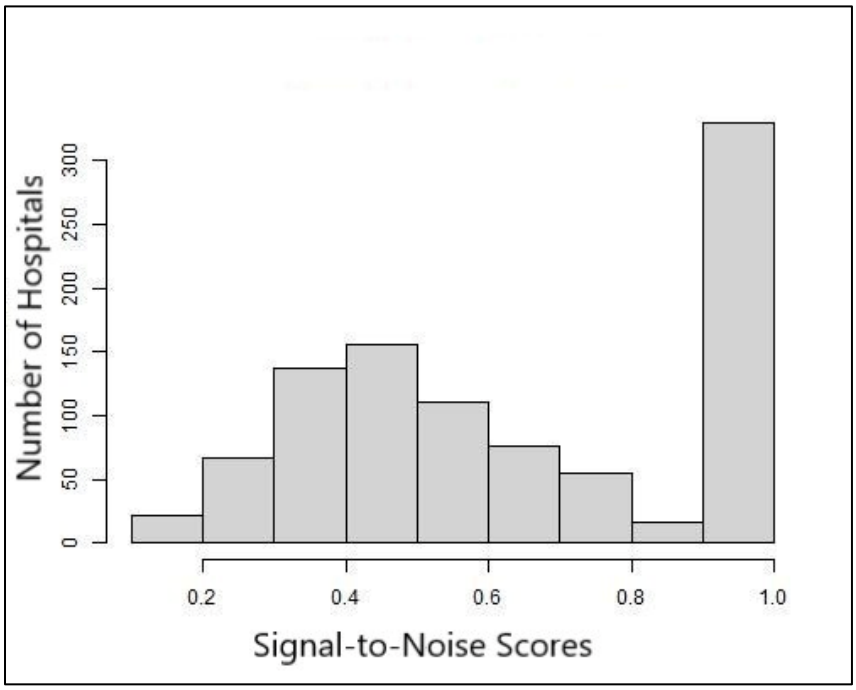



**Figure 4. Histogram of Signal-to-Noise Reliability Scores for National Hospital-Level Testing.**

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.

**Table 8. Median Reliability Scores Stratified by Number of Medicare "Benign Dizziness" Treat-and-Release Visits.**

| 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 |

***State hospital-level testing***

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

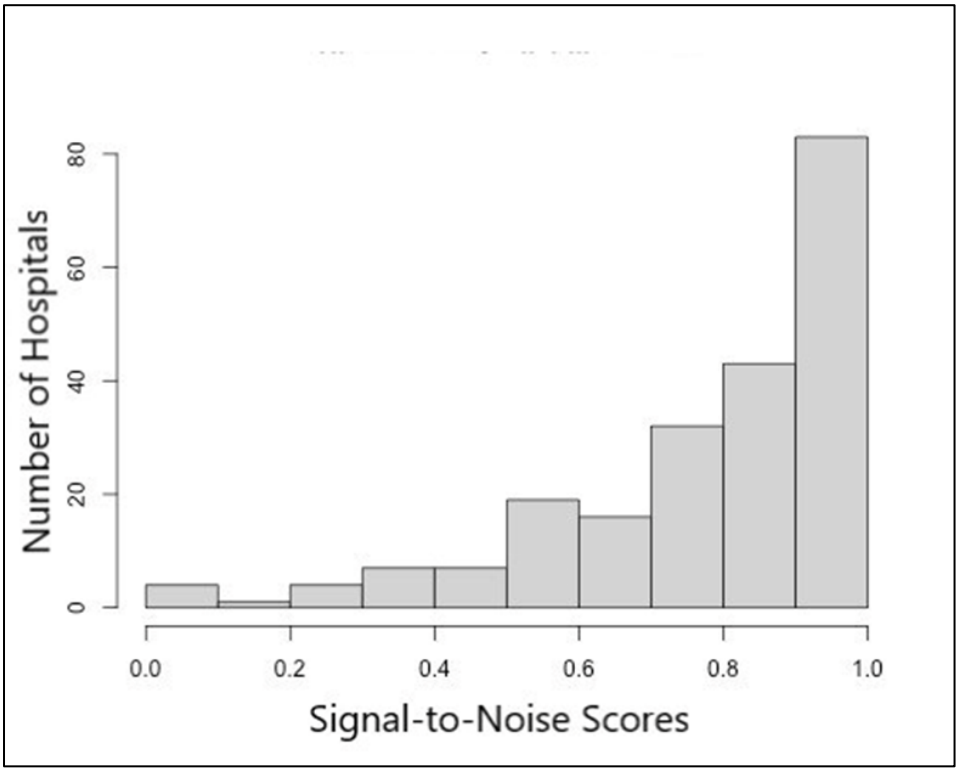



**Figure 5. Histogram of Signal-to-Noise Reliability Scores for State Hospital-Level Testing.**

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., Medicare represents only ~25% of eligible ED index visits, largely because of the age constraint [mostly patients ≥65yo]).

**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?)*

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.

## Scientific Acceptability: Validity - Testing (2b.01 - 2b.04)

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

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

☐ Accountable Entity Level (e.g., hospitals, clinicians)

☐ Empirical validity testing of the measure score

☐ Systematic assessment of face validity of performance measure score as an indicator of quality or resource use (i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance)

**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.*

***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.[56] 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.[56] in their study have a strong overlap with the ICD codes that this measure's specifications are based on.

2. McCormick et al.[57] 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[58] 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 = (true positives)/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 = (true negatives)/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.

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

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

***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,[56] 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.

2. The McCormick systematic review[57] included 77 published manuscripts between 1976–2015. The sensitivity of ICD-9 430-438/ICD-10 I60-I69 for any cerebrovascular disease was ≥82% in most [≥50%] studies, and specificity and NPV were both ≥95%. The PPV of these codes for any cerebrovascular disease was ≥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 ≤68%. In at least 50% of studies, PPVs were ≥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).

3. The Kokotailo and Hill study[58] 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.

**Table 9. Positive Predictive Values for "Benign Dizziness" Discharge Diagnosis.**

| 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% |
| TOTAL | 1826 | 1372/1372 | 100% | 99.89-100% |

** These charts were not manually reviewed but were matched based on an Epic-recorded dizziness chief complaint.*
*** Cells intentionally left empty*

**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 Table 10.

**Table 10. Negative Predictive Values for "Benign Dizziness" Discharge Diagnosis.**

| 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%* |

** 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.*
*** Cells intentionally left empty*

***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.

**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?)**

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

Both the Tirschwell[56] and the McCormick[57] 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[58] 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.

## Scientific Acceptability: Validity - Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) (2b.05 - 2b.14)

**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.*

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.

**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.*

***National hospital-level testing***

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

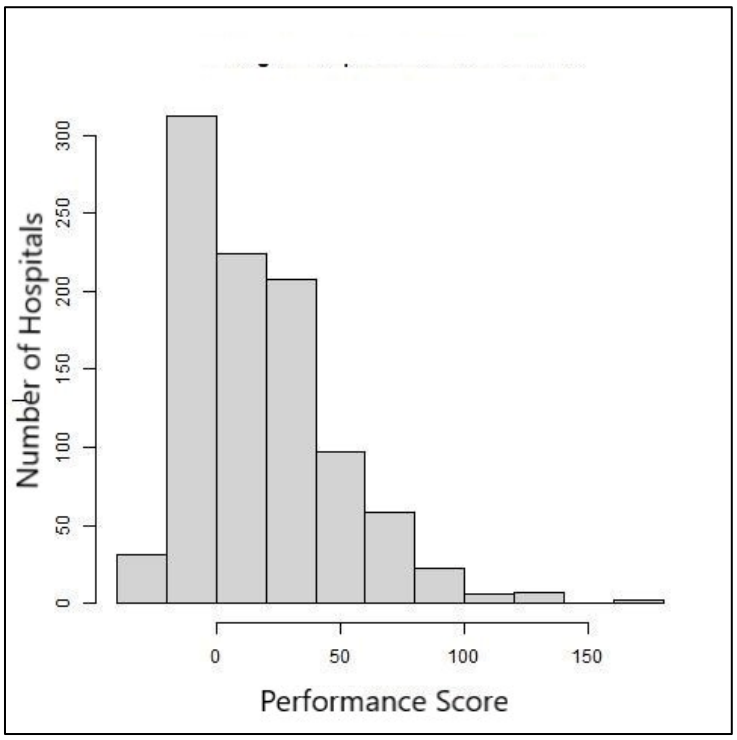



**Figure 6. Histogram of Performance Scores for National Hospital-Level Testing.**

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

- Mean: 17.70
- Median: 13.33
- 25th Percentile: -7.32
- 75th 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.

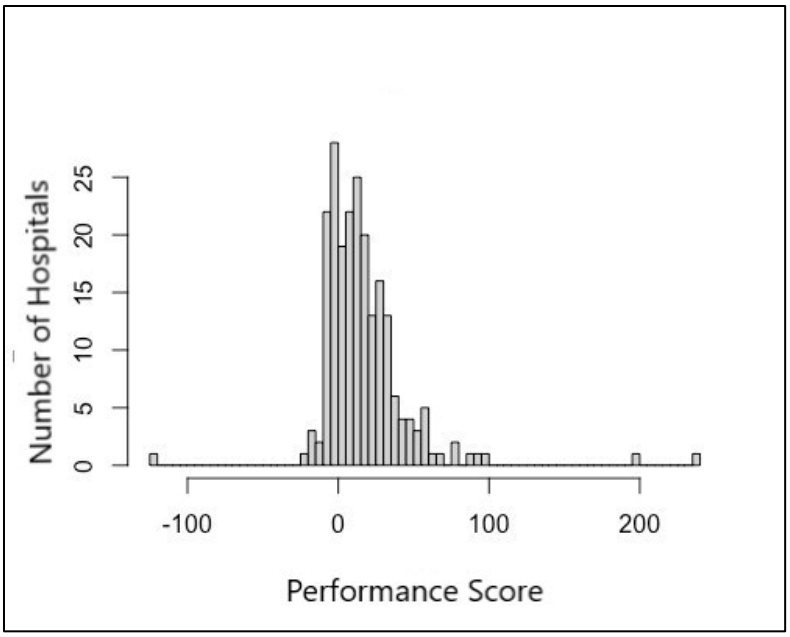



**Figure 7. Histogram of Performance Scores for State Hospital Level Testing.**

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

- Mean: 16.81
- Median: 11.27
- 25th Percentile: 0
- 75th 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)

**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?*

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$^{th}$ percentile is close to the difference between the median and the 75$^{th}$ 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

**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.*

***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:

**Table 11. Scenarios Tested as Part of the Sensitivity Analyses for Missing Stroke Admissions (Short-Term and Long-Term).**

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

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.

**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).*

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

**Table 12. Results of the Sensitivity Analyses Scenarios for Missing Stroke Admissions (Short-Term and Long-**

**Term).**

| | 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 |

**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.*

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.

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

☐ Yes, there is more than one set of specifications for this measure
☒ No, there is only one set of specifications for this measure

**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.*

Not applicable.

**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.*

Not applicable.

**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.*

Not applicable.

## Scientific Acceptability: Validity - Other Threats to Validity (Exclusions, Risk Adjustment) (2b.15 - 2b.32)

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

☒ N/A or no exclusions
☐ Yes, the measure uses exclusions.

**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?*

Not applicable.

**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.*

Not applicable.

**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.*

Not applicable.

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

☐ Statistical risk model with risk factors (specify number of risk factors)
☐ Stratification by risk category (specify number of categories)
☐ Other (please specify here: )
☒ No risk adjustment or stratification

**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.**

Not applicable.

**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.**

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 patient 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. 8). This curve matches the natural history/biological profile of major stroke following minor stroke and TIA (Fig. 9/10).

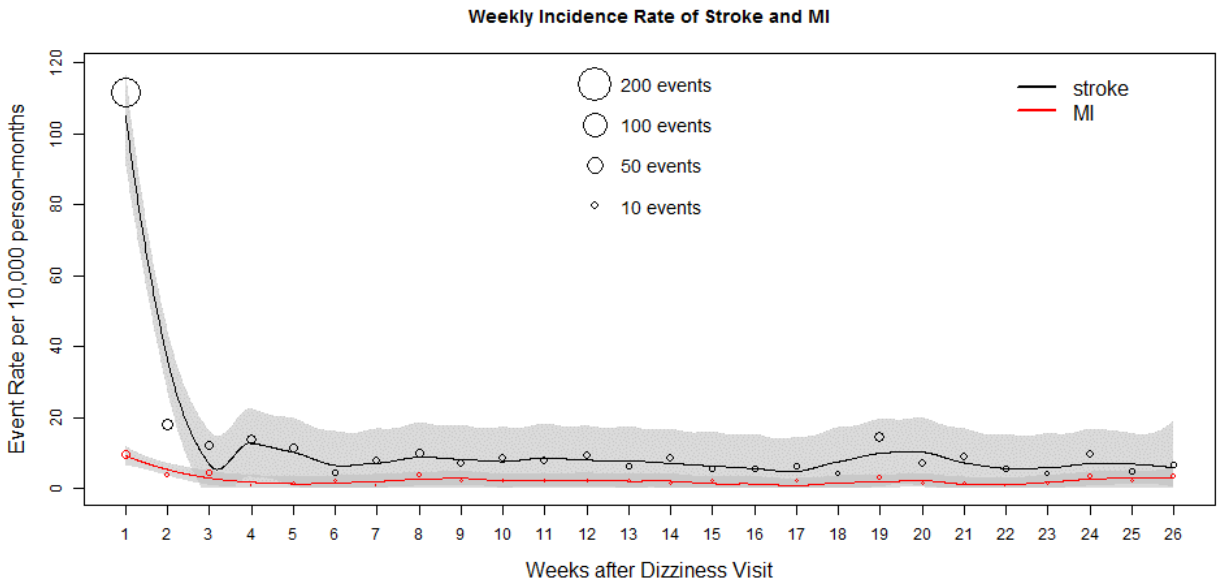



**Figure 8. Weekly incidence rate curve of stroke hospitalizations post ED treat-and-release discharge as "benign dizziness."** Kaiser Permanente Mid-Atlantic data from the performance period from 2010-2014 at all outpatient sites (ED, ambulatory care). Data reflect 56,746 treat-and-release visits for "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).

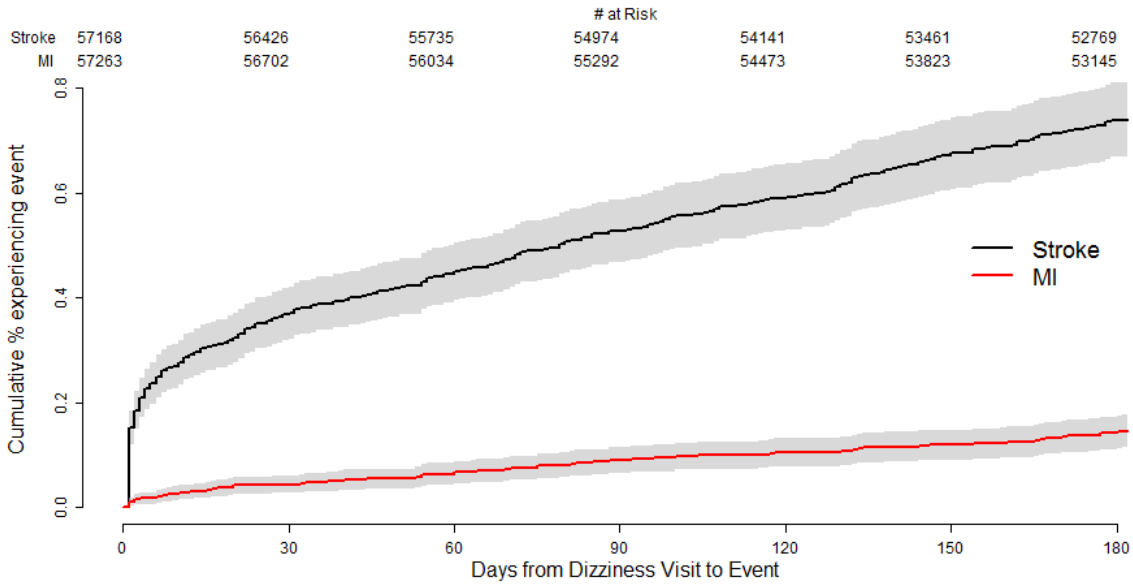



**Figure 9. Cumulative incidence curve of stroke hospitalizations post ED treat-and-release discharge**

**as "benign dizziness."** Represented here are the same data as shown in Figure 8. These data are presented here as a cumulative incidence curve for comparison to Figure 10, which illustrates the disease natural history of major stroke after transient ischemic attack (TIA) or minor stroke.

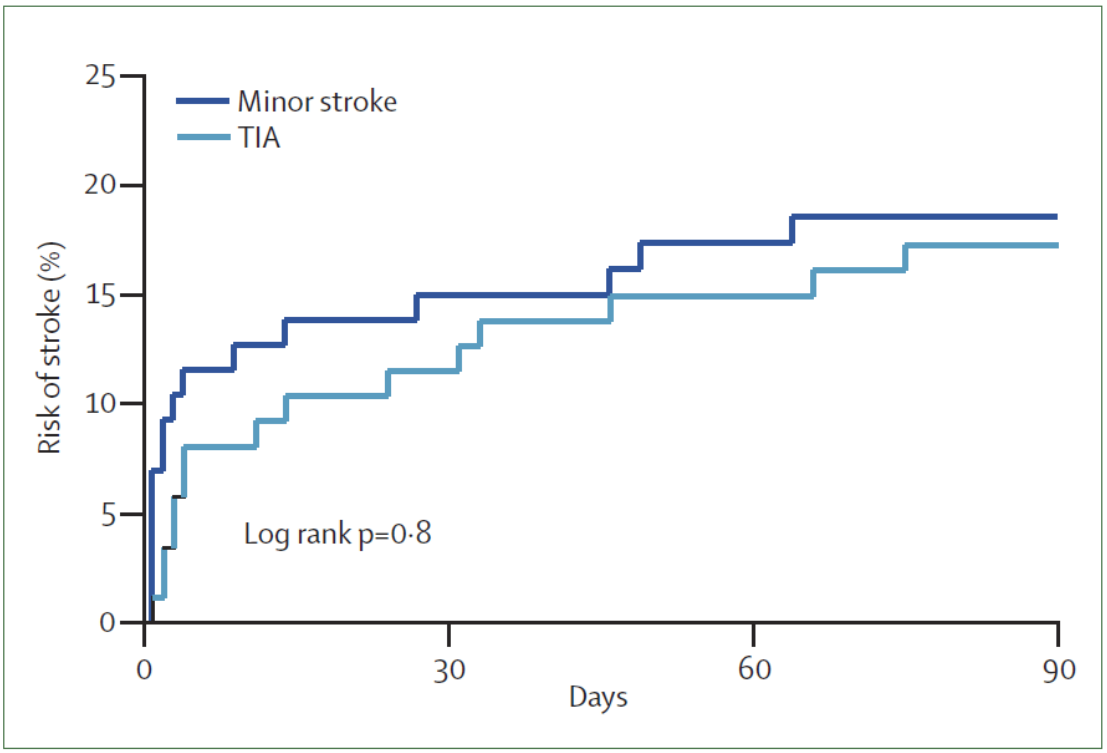



**Figure 10. Cumulative incidence curve for major stroke following TIA or minor stroke.** Data are from the Oxford Vascular Study as represented in Rothwell, Buchan, & Johnston.[21] This natural history curve matches the empirical pattern of stroke hospitalizations when some patients are diagnosed (erroneously) as "benign dizziness" and discharged home.

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 9). 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[47]).

***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%)[44]; 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.

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

☐ Published literature
☐ Internal data analysis
☐ Other (please specify here: )

Not applicable.

**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).*

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 that may place minorities at higher risk of being misdiagnosed[47]).

**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.**

Not applicable.

**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.*

Not applicable.

**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.*

Not applicable.

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

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

Not applicable.

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

Not applicable.

**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.*

Not applicable.

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

Not applicable.

**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?*

Not applicable.

**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.*

Not applicable.

## Feasibility (3.01 - 3.07)

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

☐ Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
☒ Coded by someone other than person obtaining original information (e.g., DRG, ICD-10 codes on claims)
☐ Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
☐ Other (Please describe)

**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. ALL data elements are in defined fields in electronic health records (EHRs)*

☒ ALL data elements are in defined fields in electronic claims
☐ ALL data elements are in defined fields in electronic clinical data (e.g., clinical registry, nursing home MDS, home health OASIS)
☐ ALL data elements are in defined fields in a combination of electronic sources
☐ Some data elements are in defined fields in electronic sources
☐ No data elements are in defined fields in electronic sources
☐ Patient/family reported information (may be electronic or paper)

**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.**

Not applicable.

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

Not needed, as measure is entirely claims based.

**3.05) Complete and attach the eCQM-Feasibility-Scorecard.xls file.**

Not applicable.

**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.**

The measure requires very few data elements in order to be calculated, all of which are routinely collected in the course of clinical care – discharge diagnosis codes (ICD-10-CM) and dates for emergency department (ED) visits and inpatient hospital stays. For local quality improvement purposes, these data can be gathered by institutions with little or no effort. For cross-institutional benchmarking purposes, data sets such as Medicare or AHRQ's HCUP SID and SEDD can be used. As presented in this application, the measure was calculated using both Medicare claims data and state-level SID and SEDD data.

For local quality improvement purposes, the measure can be tracked over time using only individual hospital claims data as the source. However, since patients can (and do) cross over between hospitals (i.e., discharged from ED at hospital A with "benign dizziness" and admitted for stroke to hospital B), the ideal data set would include patient follow-up across hospitals. Such follow-up is usually available when payer data are used, so optimal data sets for cross-institutional benchmarking at a national level would be those drawn from national claims data sets such as Medicare. However, because short-term cross-hospital stroke events rarely occur outside a defined geographic region, cross-institutional benchmarking can also occur at the regional level using regional health information exchanges or at the state level using curated data sets such as AHRQ's HCUP SID and SEDD data, for states where linkable data sets are available (at least 14 states currently have such capabilities[59]).

The main tradeoff when using Medicare data for national benchmarking is that Medicare data represent only ~20% of the sample of patients in any given ED. This necessarily reduces the measure's precision substantially, limiting its use to larger EDs. Also, Medicare data are restricted to older patients, so any variation in diagnostic performance based on patient age will not be detectable. An ideal data source would be a national all-payer claims database that included all ages. Until such a data source becomes readily available, however, tradeoffs are inevitable. As has been done for other measures used by CMS, a hybrid solution can be deployed if Medicare data are ultimately used for benchmarking. Specifically, hospitals with sufficient visit volumes or event rates to yield a precise result can be directly compared, while those too small for a precise result can be given individualized institutional feedback without public reporting. Such individualized results can be used for local quality improvement.

The measure, as currently defined, uses stroke returns to any hospital for the numerator definition. This definition provides the most encompassing capture of stroke hospitalizations but requires an entity like a health plan to calculate the measure, as they have access to claims from wherever the patient sought care. This choice of definition means that an individual hospital which calculates their own performance on the measure will necessarily underestimate the diagnostic adverse event rate (i.e., 30-day stroke hospitalizations), which will give a falsely better performance than occurred in reality. It is likely that for tracking diagnostic quality and safety over time within that institution, this would not matter much. However, it is even possible that the biasing effect of such data missingness might have a limited impact on cross-institutional benchmarking.

As part of our sensitivity analyses, we explored the impact of restricting the numerator definition to stroke hospitalizations only at the hospital where the patient was seen in the ED

and discharged. We found that while a hospital's calculated rate changes with this numerator restriction, a hospital's performance on the measure, relative to its peers, changes relatively little. Restricting the numerator to only same hospital strokes, we found that 81% of hospitals would either be in the same decile of performance or move just one or two deciles up or down. This supports the notion that a surrogate measure may be a meaningful way for hospitals to track their own internal performance, while their official performance is calculated from stroke returns to all hospitals from a more uniform data set. It also suggests that a combination of self-reported institutional data (with adjustment for hospital crossover rates using Medicare claims, as is done, for example, by the Maryland Hospital Rate Setting Commission) could provide a reasonable surrogate even for high-stakes public reporting or payment incentives. As a result, we believe that the measure, if endorsed by Battelle and adopted by CMS, could eventually be applied to the vast majority of hospitals through this sort of hybrid data sourcing and crossover adjustment, which would allow ~5-fold greater precision than that seen with Medicare data alone.

*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.*

There are no explicit fees or licenses associated with calculating this measure. Outside of acquiring the claims datasets themselves, all of the information needed to calculate the measure (i.e., the measure specifications, calculation algorithms, risk adjustment approach) are freely available in the public domain.

## Use (4a.01 – 4a.10)

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.

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**

☐ Public Reporting
☐ Public Health/Disease Surveillance
☐ Payment Program
☐ Regulatory and Accreditation Programs
☐ Professional Certification or Recognition Program
☐ Quality Improvement with Benchmarking (external benchmarking to multiple organizations)
☐ Quality Improvement (Internal to the specific organization)
☒ Not in use
☐ Use unknown
☐ Other (please specify here: )

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

☒ Public reporting
☒ Public Health/Disease Surveillance
☒ Payment Program
☐ Regulatory and Accreditation Program
☐ Professional Certification or Recognition Program
☒ Quality Improvement with Benchmarking (external benchmarking to multiple organizations)
☒ Quality Improvement (internal to the specific organization)
☐ Measure Currently in Use
☐ Other (please specify here: )

**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?*

This is a newly developed measure, so it is currently not being publicly reported or being used in an existing accountability program.

**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.*

*Internal Quality Improvement*
As discussed in section 3c.1, an adapted version of the measure could initially be used by hospitals for their own internal QI efforts. As a sensitivity analysis, we explored the impact of restricting the numerator definition to stroke hospitalizations only at the hospital where the patient was seen in the ED and discharged, which would allow hospitals to self-calculate their own performance on the measure. We found that while a hospital's calculated rate changes with this numerator restriction, a hospital's relative performance on the measure, relative to its peers, changes little. Restricting the numerator to only same hospital strokes, we found that 81% of hospitals would be either in the same decile of performance or move just one decile up or down. It is likely that, absent major shifts over time in local ED and inpatient visit dynamics (e.g., new hospital opening or old hospital closure), a hospital could use its own data to track performance over time without difficulty. Because of greater institution-level measure precision than is reported here (i.e., we used Medicare data which represent only about ~20% of the actual ED dizziness visits at any given hospital), even relatively small hospital EDs could track performance using a 3-year rolling window. We estimate that all but those EDs with annual volumes less than ~15,000-20,000 visits per year could do so reliably.

This approach could occur immediately for any individual hospital on a voluntary basis. Following endorsement by Battelle, such an approach could be further promoted by organizations such as the Society to Improve Diagnosis in Medicine (https://www.improvediagnosis.org/) and the multi-stakeholder Coalition to Improve Diagnosis, which currently has more than 60 partner organizations (https://www.improvediagnosis.org/coalition/). Adoption by hundreds of hospitals could potentially happen within 12-18 months of a Battelle measure endorsement.

*Public Health/Disease Surveillance*
The measure lends itself to having a federal agency, such as the Agency for Healthcare Research and Quality (AHRQ), calculate aggregated hospital performance using a national dataset (e.g., HCUP dataset) and track national performance on the measure over time. There would also be the opportunity to stratify aggregated national performance by key patient

sociodemographic variables (e.g., race, gender, age) and report out those findings through their annual national disparities report. HCUP data have already been used to address the issue of misdiagnosing dizziness and stroke, so this sort of work could be reasonably be incorporated within 1-3 years of a Battelle measure endorsement.

*Public Reporting/External Benchmarking*
Public reporting and external benchmarking initially on a voluntary basis could occur through the Leapfrog Group. Participating hospitals could self-report data on all of their patients, and an adjustment for estimated crossover fractions could be made based upon payer claims data analysis (public or commercial) [see 3c.1], through partnership between Leapfrog and relevant payers participating in Leapfrog's Value-Based Purchasing program. This sort of program could potentially be implemented within 2-4 years of a Battelle measure endorsement.

*Payment Program*
While public reporting of the measure would definitely need to precede the use of the measure in a payment program, we would anticipate that the measure could be incorporated into hospital pay-for-performance programs, with possible adoption by the Centers for Medicare and Medicaid Services (CMS) and other payers. For example, ED patients with dizziness could be covered by a symptom-related overall payment in the ED (e.g., $1,000 for a diagnostic evaluation for dizziness, to include all usual care fees, imaging, and consultations); then institutions could be held accountable to diagnostic accuracy (e.g., this measure was used to produce a penalty for those institutions who missed more strokes than their peers and a bonus for those who missed fewer). This sort of program could potentially be implemented within 4-6 years of a Battelle measure endorsement.

**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.*

The measure is being implemented at Johns Hopkins as a diagnostic outcome metric in our stroke misdiagnosis reduction initiative through the Armstrong Institute Center for Diagnostic Excellence. It has already been incorporated into an operational diagnostic performance dashboard at Kaiser Permanente, Mid-Atlantic States (KPMAS), with whom Johns Hopkins (the measure steward) has been collaborating. An initial version of the dashboard co-developed by the two institutions was described in a 2018 publication.[48]

**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.**

The measure is being reported to ED quality and safety leaders and the Director of the Armstrong Institute for Patient Safety and Quality at Johns Hopkins (who is also Sr. VP, Patient Safety and Quality for Johns Hopkins Medicine) on an annual basis, as recommended for the current measure parameterization (3-year rolling window updated annually). Data from

within Johns Hopkins Health System (5 adult EDs), plus non-JHHS stroke admissions (out-of-network crossovers admitted to other hospitals, such as University of Maryland) are included. The latter are accessed via the state-designated regional health information exchange (HIE) for Maryland known as CRISP (https://crisphealth.org/), with whom we have established an ongoing partnership with quarterly updates to the data warehouse for the measure. Using this approach, the measure could readily be deployed throughout Maryland if endorsed by Battelle. Measures, trends, and incidence rate curves are provided to patient safety leaders. Education and explanation about both the methods and interpretation of findings occur during annual strategic planning meetings of the Patient and Family Centered Care (PFCC) committee which includes patient safety. The ED's Associate Medical Director for Patient Safety and Quality, who is heavily engaged in the measurement work also briefs other members of the ED leadership team (e.g., Chairman, Medical Director, Research Director).

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

Based on these measures and the success of the Tele-Dizzy program, Johns Hopkins has agreed to extend its stroke reduction initiatives to all 5 hospitals within the Johns Hopkins Health System with adult EDs. KPMAS, as a consequence of its measurement efforts using this approach, has implemented an educational program for evaluating dizziness for its clinical faculty. KPMAS is also considering submitting grant proposals in partnership with Johns Hopkins to extend the Tele-Dizzy program to its Clinical Decision Units (CDUs), (which are similar to EDs, but do not take high-severity [Level 1] patients).

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

Feedback on the measure from ED physicians in the quality improvement space has been very positive, overall. NQF's Advancing Chief Complaint-Based Quality Measurement (final report June 24, 2019)[60] focused on ED quality measurement and included more than a dozen leaders from emergency medicine from around the US. This group deemed the "rate of missed stroke diagnosis for patients with a presenting problem of dizziness/vertigo" using the SPADE method one of just three diagnostic safety and quality measures "IMPORTANT AND FEASIBLE FOR DEVELOPMENT NOW." A recent comprehensive AHRQ report (systematic review and meta-analysis) on Diagnostic Errors in the Emergency Department also highlighted the importance of this method of measurement for quality improvement purposes and national benchmarking.[61]

We have received similar feedback from all of our ED physician partners focused on quality improvement as part of our SPADE measure development program (including partners at KPMAS, Kaiser Permanente Southern California, and the American College of Emergency Physicians as part of AHRQ R01 HS 27614: Towards a National Diagnostic Excellence Dashboard Partnering with Stakeholders to Construct Evidence Based Operational Measures of Misdiagnosis Related Harms [PI: Newman-Toker]). This stakeholder feedback as part of R01 HS 27614 is detailed above in Section **1a.02**. Briefly, the ACEP surveys found that over 80% of both groups surveyed (frontline clinicians 81%, medical directors 85%) said that

receiving hospital/ED-level feedback on missed stroke in dizziness/vertigo presentations would improve their practice and the quality of care for patients with dizziness/vertigo, and over 90% of both groups said they would welcome such feedback.

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

Feedback on the SPADE measurement approach (and specifically as it relates to stroke misdiagnosis) has been taken from multiple stakeholders since 2016 through presentations at national meetings including the Diagnostic Error in Medicine Meeting, the Diagnostic Error in Medicine Research Summit, and via multiple publications. Increasingly this measurement approach is recognized as an important tool in the diagnostic quality and safety measurement armamentarium, as articulated now in three related NQF reports which have recognized its increasing relevance[42,60,62] and a comprehensive AHRQ Evidence-based Practice Center report on Diagnostic Errors in the ED.[61]

**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.**

Feedback has led to modified use of code sets for the stroke numerator. On the basis of feedback, a modified denominator version (using a presenting symptom of dizziness, rather than a discharge diagnosis), is being developed in parallel; this is not presented here because, as yet, chief complaint data are not yet consistently reported in various public use data sets, so they cannot be readily used to support the analyses presented here.

Feedback on the need for balancing measures has been clear. Measures related to use of CT and MRI neuroimaging must be deployed in parallel with the deployment of such a measure, given concerns for diagnostic test overuse as a consequence of public reporting and accountability related to missed stroke. Such balancing measures are again readily assessed using claims data sets.

## Usability (4b.01 - 4b.03)

**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.**

A recently published study deploying a related strategy (analogous to SPADE's "look back," but not symptom-specific) and also using Medicare data suggested a trend towards slightly increased risk of stroke misdiagnosis from 2007-2014.[63] Looking across the three 3-year time periods for which our measure was calculated, we have seen small, but steady improvement over time. The mean performance on the measure has improved slightly in each successive time period (where lower performance is desirable) and the standard deviation on the measure has shrunk. Despite this apparent improvement, median hospital performance on the measure in 7 of the 10 deciles remains at or above zero, indicating there is still significant room for improvement at most hospitals.

It is possible that the discrepancy between the prior study and our Medicare data is methodological, but it is more likely that this reflects a general upward trend in overuse of MRI neuroimaging, particularly at larger hospitals (i.e., the ones included in the current analysis), rather than improvement in diagnostic acumen. This conjecture is supported by our analysis showing that larger hospitals and those obtaining more MRIs are outperforming smaller hospitals and those obtaining fewer MRIs. A recent analysis by our team of the CDC's National Hospital Ambulatory Medical Care Survey data found imaging for dizziness has continued to rise over time and outpaces the average across other ED complaints substantially. However, it is also known that imaging for dizziness diagnosis varies substantially by institution, with some community-based EDs having MRI rates of just 0.8%[64] and some large academic centers having current MRI rates of up to 20%.[10] This again reinforces the need for balance measures, as noted.

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

As yet, we have detected no unexpected findings (positive or negative) during the relatively recent and small-scale deployment of this measure, including no unintended impacts on patients.

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

Not applicable.

## Related and Competing (5.01 - 5.06)

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 endorsed. Please review and update questions 5.01, 5.02, and 5.03 accordingly.

**5.01) Search and select all endorsed related measures (conceptually, either same measure focus or target population) by going to the PQM website.**

None identified.

*(Can search and select measures.)*

**5.02) Search and select all endorsed competing measures (conceptually, the measures have both the same measure focus or target population) by going to the PQM website.**

*(Can search and select measures.)*

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

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

☐ Yes
☐ No

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

**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.*

## Additional (1 - 9)

**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.**

☒ Available in attached file
☐ No appendix
☐ Available at measure-specific web page URL identified in sp.09

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

The development work of this measure has been within Johns Hopkins. The measure has been tested with external organizations.

**3) Indicate the year the measure was first released.**

Has not been released; work started in 2017/2018.

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

May 2023

**5) Indicate the frequency of review, or an update schedule, for this measure.**
0

No set schedule has been set for this measure.

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

No set schedule has been set for this measure.

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

Not applicable.

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

Not applicable.

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

Not applicable.

**References**

1. Savitz SI, Caplan LR, Edlow JA. Pitfalls in the diagnosis of cerebellar infarction. *Acad Emerg Med*. 2007;14(1):63-68.

2. Missed Stroke Diagnosis - John Michael Night's Story. 2020. https://www.improvediagnosis.org/stories_posts/missed-stroke-diagnosis/

3. Tarnutzer AA, Berkowitz AL, Robinson KA, Hsieh YH, Newman-Toker DE. Does my dizzy patient have a stroke? A systematic review of bedside diagnosis in acute vestibular syndrome. *Can Med Assoc J*. 2011;183(9):E571-92. doi:10.1503/cmaj.100174

4. Cohn B. Can Bedside Oculomotor (HINTS) Testing Differentiate Central From Peripheral Causes of Vertigo? *Ann Emerg Med*. 2014. doi:10.1016/j.annemergmed.2014.01.010

5. Krishnan K, Bassilious K, Eriksen E, et al. Posterior circulation stroke diagnosis using HINTS in patients presenting with acute vestibular syndrome: A systematic review. *Eur Stroke J*. 2019;4(3):233-239. doi:10.1177/2396987319843701

6. Ohle R, Montpellier RA, Marchadier V, et al. Can Emergency Physicians Accurately Rule Out a Central Cause of Vertigo Using the HINTS Examination? A Systematic Review and Meta-analysis. *Acad Emerg Med*. 2020;27(9):887-896. doi:10.1111/acem.13960

7. Newman-Toker DE, Della Santina CC, Blitz AM. Vertigo and hearing loss. *Handb Clin Neurol*. 2016;136:905-921. doi:10.1016/B978-0-444-53486-6.00046-6

8. Tarnutzer AA, Lee SH, Robinson KA, Wang Z, Edlow JA, Newman-Toker DE. ED misdiagnosis of cerebrovascular events in the era of modern neuroimaging: A meta-analysis. *Neurology*. 2017;88(15):1468-1477. doi:10.1212/WNL.0000000000003814

9. Nham B, Reid N, Bein K, et al. Capturing vertigo in the emergency room: three tools to double the rate of diagnosis. *J Neurol*. 2022;269(1):294-306. doi:10.1007/s00415-021-10627-1

10. Gold D, Peterson S, McClenney A, Tourkevich R, Brune A, Choi W, Shemesh A, Maliszewski B, Bosley J, Otero-Millan J, Fanai M, Roberts D, Tevzadze N, Zee DS, Newman-Toker DE. Diagnostic impact of a device-enabled remote "Tele-Dizzy" consultation service [abstract]. *Diagnostic Error in Medicine, 12th Annual Conference*. 2019.

11. Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Arch Intern Med*. 2009;169(22):2078-2086. doi:10.1001/archinternmed.2009.427

12. Vanni S, Nazerian P, Casati C, et al. Can emergency physicians accurately and reliably assess acute vertigo in the emergency department? *Emerg Med Australas*. 2015. doi:10.1111/1742-6723.12372

13. Tarnutzer AA, Gold D, Wang Z, et al. Impact of Clinician Training Background and Stroke Location on Bedside Diagnostic Accuracy in the Acute Vestibular Syndrome - A Meta-Analysis. *Ann Neurol*. 2023. doi:10.1002/ana.26661

14. Edlow JA, Carpenter C, Akhter M, et al. Guidelines for reasonable and appropriate care in the emergency department 3 (GRACE-3): Acute dizziness and vertigo in the emergency department. *Acad Emerg Med*. 2023;30(5):442-486. doi:10.1111/acem.14728

15. Pan Y, Elm JJ, Li H, et al. Outcomes Associated With Clopidogrel-Aspirin Use in Minor Stroke or Transient Ischemic Attack: A Pooled Analysis of Clopidogrel in High-Risk Patients With Acute Non-

Disabling Cerebrovascular Events (CHANCE) and Platelet-Oriented Inhibition in New TIA and Minor Ischemic Stroke (POINT) Trials. *JAMA Neurol*. 2019;76(12):1466-1473. doi:10.1001/jamaneurol.2019.2531

16. Kuruvilla A, Bhattacharya P, Rajamani K, Chaturvedi S. Factors associated with misdiagnosis of acute stroke in young adults. *Journal of Stroke and Cerebrovascular Diseases*. 2011;20(6):523-527. doi:10.1016/j.jstrokecerebrovasdis.2010.03.005

17. Cano LM, Cardona P, Quesada H, Mora P, Rubio F. [Cerebellar infarction: prognosis and complications of vascular territories]. *Neurologia*. 2012;27(6):330-335. doi:10.1016/j.nrl.2011.12.009

18. Edlow JA, Newman-Toker DE, Savitz SI. Diagnosis and initial management of cerebellar infarction. *Lancet Neurol*. 2008;7(10):951-964.

19. Jauch EC, Saver JL, Adams Jr. HP, et al. Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2013;44(3):870-947. doi:10.1161/STR.0b013e318284056a

20. Flossmann E, Rothwell PM. Prognosis of vertebrobasilar transient ischaemic attack and minor stroke. *Brain*. 2003;126(Pt 9):1940-1954.

21. Rothwell PM, Buchan A, Johnston SC. Recent advances in management of transient ischaemic attacks and minor ischaemic strokes. *Lancet Neurol*. 2006;5(4):323-331.

22. Paul NL, Simoni M, Rothwell PM. Transient isolated brainstem symptoms preceding posterior circulation stroke: a population-based study. *Lancet Neurology*. 2013;12(1):65-71. doi:10.1016/S1474-4422(12)70299-5

23. Hacke W, Kaste M, Bluhmki E, et al. Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. *N Engl J Med*. 2008;359(13):1317-1329. doi:359/13/1317

24. Lyden P. Thrombolytic therapy for acute stroke--not a moment to lose. *N Engl J Med*. 2008;359(13):1393-1395. doi:359/13/1393

25. Lavallee PC, Meseguer E, Abboud H, et al. A transient ischaemic attack clinic with round-the-clock access (SOS-TIA): feasibility and effects. *Lancet Neurol*. 2007;6(11):953-960. doi:S1474-4422(07)70248-X

26. Rothwell PM, Giles MF, Chandratheva A, et al. Effect of urgent treatment of transient ischaemic attack and minor stroke on early recurrent stroke (EXPRESS study): a prospective population-based sequential comparison. *Lancet*. 2007;370(9596):1432-1442. doi:S0140-6736(07)61448-2

27. Bhattacharyya N, Gubbels SP, Schwartz SR, et al. Clinical Practice Guideline: Benign Paroxysmal Positional Vertigo (Update). *Otolaryngol Head Neck Surg*. 2017;156(3_suppl):S1-S47. doi:10.1177/0194599816689667

28. Fife TD, Iverson DJ, Lempert T, et al. Practice parameter: therapies for benign paroxysmal positional vertigo (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*. 2008;70(22):2067-2074.

29. Hilton MP, Pinder DK. The Epley (canalith repositioning) manoeuvre for benign paroxysmal positional vertigo. *Cochrane Database of Systematic Reviews*. 2014;(12).

doi:10.1002/14651858.CD003162.pub3

30. van den Broek EM, van der Zaag-Loonen HJ, Bruintjes TD. Systematic Review: Efficacy of Gufoni Maneuver for Treatment of Lateral Canal Benign Paroxysmal Positional Vertigo with Geotropic Nystagmus. *Otolaryngol Head Neck Surg*. 2014;150(6):933-938. doi:10.1177/0194599814525919

31. Lopez-Escamez JA, Gamiz MJ, Fernandez-Perez A, Gomez-Finana M, Sanchez-Canet I. Impact of treatment on health-related quality of life in patients with posterior canal benign paroxysmal positional vertigo. *Otology & Neurotology*. 2003;24(4):637-641.

32. Hillier S, McDonnell M. Is vestibular rehabilitation effective in improving dizziness and function after unilateral peripheral vestibular hypofunction? An abridged version of a Cochrane Review. *Eur J Phys Rehabil Med*. 2016;52(4):541-556.

33. Strupp M, Zingler VC, Arbusow V, et al. Methylprednisolone, valacyclovir, or the combination for vestibular neuritis. *NEnglJMed*. 2004;351(4):354-361.

34. Sacco AY, Self QR, Worswick EL, et al. Patients' Perspectives of Diagnostic Error: A Qualitative Study. *J Patient Saf*. 2020. doi:10.1097/PTS.0000000000000642

35. Newman-Toker DE, Edlow JA. TiTrATE: A Novel, Evidence-Based Approach to Diagnosing Acute Dizziness and Vertigo. *Neurol Clin*. 2015;33(3):577-599, viii. doi:10.1016/j.ncl.2015.04.011

36. Eagles D, Stiell IG, Clement CM, et al. International survey of emergency physicians' priorities for clinical decision rules. *Acad Emerg Med*. 2008;15(2):177-182. doi:10.1111/j.1553-2712.2008.00035.x

37. Kene M V, Ballard DW, Vinson DR, Rauchwerger AS, Iskin HR, Kim AS. Emergency Physician Attitudes, Preferences, and Risk Tolerance for Stroke as a Potential Cause of Dizziness Symptoms. *West J Emerg Med*. 2015;16(5):768-776. doi:10.5811/westjem.2015.7.26158

38. Kerber KA, Newman-Toker DE. Misdiagnosing Dizzy Patients: Common Pitfalls in Clinical Practice. *Neurol Clin*. 2015;33(3):565-575, viii. doi:10.1016/j.ncl.2015.04.009

39. Grewal K, Austin PC, Kapral MK, Lu H, Atzema CL. Missed strokes using computed tomography imaging in patients with vertigo: population-based cohort study. *Stroke*. 2015;46(1):108-113. doi:10.1161/strokeaha.114.007087

40. Committee on Diagnostic Error in Health Care; Board on Health Care Services; Institute of Medicine; The National Academies of Sciences, Engineering and Medicine. *Improving Diagnosis in Health Care*. (Balogh EP, Miller BT, Ball JR, eds.). National Academies Press; 2015. doi:10.17226/21794

41. Henriksen K, Dymek C, Harrison MI, Brady PJ, Arnold SB. Challenges and opportunities from the Agency for Healthcare Research and Quality (AHRQ) research summit on improving diagnosis: a proceedings review. *Diagnosis (Berl)*. 2017;4(2):57-66. doi:10.1515/dx-2017-0016

42. National Quality Forum. *Improving Diagnostic Quality and Safety*. 2017. https://www.qualityforum.org/Projects/i-m/Improving_Diagnostic_Accuracy/Final_Report.aspx

43. Newman-Toker DE, Tucker L, Berenson R, et.al. The Roadmap for Research to Improve Diagnosis, Part 1: Converting National Academy of Medicine Recommendations into Policy Action. 2018:11. https://www.improvediagnosis.org/wp-content/uploads/2018/10/policy_roadmap_for_diagnosti.pdf

44. Newman-Toker DE. Missed stroke in acute vertigo and dizziness: It is time for action, not debate. *Ann*

*Neurol*. 2016;79(1):27-31. doi:10.1002/ana.24532

45. Newman-Toker DE, Pronovost PJ. Diagnostic errors--the next frontier for patient safety. *JAMA*. 2009;301(10):1060-1062. doi:10.1001/jama.2009.249

46. Bery A, Chang T, Wang Z, Chuang H, Newman-Toker D. Stroke risk after outpatient diagnosis of benign vertigo varied across specialties. *Tzu Chi Med J*. 2017;29(5 Supplement 1):S21-S22.

47. Newman-Toker DE, Moy E, Valente E, Coffey R, Hines AL. Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample. *Diagnosis (Berl)*. 2014;1(2):155-166. doi:10.1515/dx-2013-0038

48. Mane KK, Rubenstein KB, Nassery N, et al. Diagnostic performance dashboards: tracking diagnostic errors using big data. *BMJ Qual Saf*. 2018;27(7):567-570. doi:10.1136/bmjqs-2018-007945

49. von Kleist T, Mohammed A, Chokhawala H, Vidal G, Zweifler R. Gender Disparities in Misdiagnosis of Ischemic Stroke or TIA During Telestroke Consultation (P2.3-018). *Neurology*. 2019;92(15 Supplement):P2.3-018. http://n.neurology.org/content/92/15_Supplement/P2.3-018.abstract

50. Fast facts on U.S. hospitals, 2020. *American Hospital Association*. 2020. https://www.aha.org/statistics/fast-facts-us-hospitals

51. Muelleman RL, Sullivan AF, Espinola JA, Ginde AA, Wadman MC, Camargo CA. Distribution of emergency departments according to annual visit volume and urban-rural status: implications for access and staffing. *Acad Emerg Med*. 2010;17(12):1390-1397. doi:10.1111/j.1553-2712.2010.00924.x

52. Individual hospital statistics for Florida. *American Hospital Directory*. 2023. https://www.ahd.com/states/hospital_FL.html

53. Newman-Toker DE, Hsieh YH, Camargo CAJ, Pelletier AJ, Butchy GT, Edlow JA. Spectrum of dizziness visits to US emergency departments: cross-sectional analysis from a nationally representative sample. *Mayo Clin Proc*. 2008;83(7):765-775. doi:10.4065/83.7.765

54. Kilduff L. Which U.S. states have the oldest populations? *Population Reference Bureau*. 2021. https://www.prb.org/resources/which-us-states-are-the-oldest/

55. Adams JL. *The Reliability of Provider Profiling: A Tutorial*. RAND Corporation; 2009. doi:10.7249/tr653

56. Tirschwell DL, Longstreth WT. Validating administrative data in stroke research. *Stroke*. 2002;33(10):2465-2470.

57. McCormick N, Bhole V, Lacaille D, Avina-Zubieta JA. Validity of Diagnostic Codes for Acute Stroke in Administrative Databases: A Systematic Review. *PLoS One*. 2015;10(8):e0135834. doi:10.1371/journal.pone.0135834

58. Kokotailo RA, Hill MD. Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10. *Stroke*. 2005;36(8):1776-1781. doi:10.1161/01.STR.0000174293.17959.a1

59. Metcalfe D, Zogg CK, Haut ER, Pawlik TM, Haider AH, Perry DC. Data resource profile: State Inpatient Databases. *Int J Epidemiol*. 2019;48(6):1742-1742h. doi:10.1093/ije/dyz117

60. National Quality Forum. *Advancing Chief Complaint-Based Quality Measurement [Final Report]*. 2019. https://www.qualityforum.org/Publications/2019/06/Advancing_Chief_Complaint-Based_Quality_Measurement_Final_Report.aspx

61. Newman-Toker DE, Peterson SM, Badihian S, et al. *Diagnostic Errors in the Emergency Department: A Systematic Review*. 2022. doi:10.23970/AHRQEPCCER258

62. National Quality Forum. *Improving Diagnostic Quality and Safety/Reducing Diagnostic Error: Measurement Considerations [Final Report]*. 2020. https://www.qualityforum.org/Publications/2020/10/Reducing_Diagnostic_Error__Measurement_Considerations_-_Final_Report.aspx

63. Waxman DA, Kanzaria HK, Schriger DL. Unrecognized Cardiovascular Emergencies Among Medicare Patients. *JAMA Intern Med*. 2018;178(4):477-484. doi:10.1001/jamainternmed.2017.8628

64. Kim AS, Sidney S, Klingman JG, Johnston SC. Practice variation in neuroimaging to evaluate dizziness in the ED. *Am J Emerg Med*. 2012;30(5):665-672. doi:10.1016/j.ajem.2011.02.038