NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3614

Corresponding Measures:

De.2. Measure Title: Hospitalization After Release with Missed Dizzy Stroke (H.A.R.M Dizzy-Stroke)

Co.1.1. Measure Steward: Johns Hopkins Armstrong Institute for Patient Safety and Quality

De.3. Brief Description of Measure: This outcome measure tracks the rate of patients admitted to the hospital for a stroke within 30 days of being treated and released from the 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"). The measure accounts for the epidemiologic base rate of stroke in the population under study using a risk difference approach (observed [short-term rate] minus expected [long-term rate]).

1b.1. Developer Rationale: Diagnostic error is a major public health problem.[1] The lack of operational measures is a critical barrier to improving diagnosis.[2,3] 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.[4] Among vascular events, missed stroke is the leading cause of serious harm to patients. Misdiagnosis of stroke disproportionately occurs when symptoms and signs are not typical or obvious.[5,6] Among strokes, the most commonly missed clinical presentation is patients presenting dizziness or vertigo, easily mistaken for inner ear disease.[5] In US emergency departments (ED) each year, an estimated 45,000-75,000 patients with strokes presenting dizziness or vertigo are missed and erroneously discharged.[6]

ED patients with acute dizziness and vertigo could be diagnosed correctly using evidence-based bedside examinations,[7,8] but there is currently a large evidence-practice gap[9] in ED diagnosis, resulting in substantial harms to patients.[6] Without timely, accurate diagnosis, these patients suffer misdiagnosis-related harms[10] from lack of prompt treatment.[5] The most common harm is preventable major stroke after minor stroke or transient ischemic attack (TIA),[11,12] with major stroke leading to subsequent hospitalization. Crude short-term stroke hospitalization rates per 10,000 dizziness discharges from the ED vary at least from 20-80.[6] Adjusting for baseline stroke risk across groups does not eliminate practice variation.[13]

This outcome measure tracks the rate of patients admitted to the hospital for a stroke within 30 days of being treated and released from the 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"). This measure is the first operationally-viable performance measure of stroke misdiagnosis for the hospital setting. Hospital EDs will be able to use the measure internally to track their performance over time, as they work to implement interventions to reduce misdiagnosis of strokes. The measure can also be used by external entities for public reporting and pay-for-performance, as external pressures to encourage hospital improvements.

S.4. Numerator Statement: The number of ED index visits during the performance period that are followed within 30 days by an inpatient hospital admission to any hospital that ends in a primary diagnosis of stroke.

S.6. Denominator Statement: Patients discharged from the ED with "benign dizziness" as the primary diagnosis code, counting a patient's first such discharge during the performance period (an "index visit) and all subsequent such discharges that fall outside a 360-day follow-up window from the previous qualifying "index visit".

S.8. Denominator Exclusions: The measure has no exclusions. All patients discharged from the ED with "benign dizziness" as their primary diagnosis code are included in the measure denominator.

De.1. Measure Type: Outcome

S.17. Data Source: Claims

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

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

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

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?

Preliminary Analysis: New Measure

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary

- Brief background: This is intended to be a measure of patients who had a treat-and-release ED visit with a diagnosis of "benign dizziness" who were discharged and later had a stroke, with the suggestion being that the dizziness treat-and-release ED visit reflected a potentially missed stroke diagnosis. Specifically, it is an outcome measure of the number of ED index visits discharged during the performance period followed within 30 days by an inpatient hospital admission to any hospital with a primary diagnosis of stroke.

- Evidence suggesting patient value and meaningfulness include:
    - Dizziness is commonly misdiagnosed in the ED, both with respect to stroke and inner ear disorders; misdiagnosis rates as high as 80% have been documented in the literature.
    - Patients hospitalized for stroke (n~190 000 admissions from 9 US states in 2009) are more likely to have had a treat-and-release ED visit for so-called 'benign' dizziness within the prior 14 days than have had an ED visit for a different chief complaint.
    - 'Benign' dizziness treat-and-release discharges from the ED (n~30 000 visits per year) are more likely to return for an inpatient stroke admission within the subsequent 30 days than a heart attack admission.
- Evidence demonstrating relationship between outcome and healthcare structure, process, intervention or service include:
    - Developer cites several studies 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.

Question for the Standing Committee:

- *Does the Standing Committee agree that the evidence demonstrates interventions that providers can implement towards the improvement of patient outcomes and measure performance?*

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

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

Maintenance measures – increased emphasis on gap and variation

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

(1/1/2015-12/31/2017 – note this is an overall 3 year window); 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)

(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)

(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)

Disparities

- *Differences by Gender*
    - 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. Females were more likely than males to be given an "uncertain" diagnosis. Misdiagnoses were lower among men (OR 0.75) than women.
- *Differences by Age*
    - Patient's age <=35 years (P=.05) were more likely to be misdiagnosed.
    - 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).
- *Differences by Race*
    - 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

Questions for the Standing Committee:

- *Does the Standing Committee agree that there a gap in care that warrants a national performance measure?*
- *Based on the disparities data, does the Standing Committee think there should be risk adjustment by socioeconomic status (SES) or other data*? Note the discussion of this by the Scientific Methods Panel (SMP) is paraphrased below.

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

Committee Pre-evaluation Comments:

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure."

- I think that evidence showing that there are reliable tests that can be done at bedside that will detect strokes are a little lacking. Only the VOG data support this, but this is not widely available. It is not clear that there is more careful exam, etc. that will improve performance on this measure.
- yes
- Strong evidence supporting that better neurological assessment can improve this measure
- If the measure is intended to identify the rate of missed posterior circulation acute ischemic events, this is a clinical issue and challenge for physicians in general. Despite the claim that acute dizziness and vertigo could be diagnosed correctly using EBM (likely the same tool developed by the measure authors), it is not demonstrated to be generalizable or sufficiently effective in routine practice. Even in the hands of experienced practitioners, definitive diagnosis is often only possible with advanced cerebrovascular imaging, namely MRI.
- Evidence supports a relationship between previous treat/release for dizziness and subsequent hospitalization for stroke. If dizziness is a common symptom in the target population - and potentially for other causes - the relationship between the two events may not be as strong as the developers suggest

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- There is a disparities gap.
- some concern
- Disparities for females in terms of stroke diagnosis
- There is variability in care and often reflects local issues of physician expertise, access to neurologic consultation, and advanced imaging.
- The developer suggests that the measure will highlight the extent to which a facility may have missed an opportunity to identify a stroke at first presentation (i.e., at benign dizziness). It is unclear what change in procedures might be adopted, without certainty which cases were misdiagnoses

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data

2c. For composite measures: empirical analysis support composite approach

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators: NQF Scientific Methods Panel Subgroup 2

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

This measure was reviewed by the Scientific Methods Panel (SMP) and discussed on the call. A summary of the measure and the Panel discussion is provided below.

Description: This is a new outcome measure tracks the rate of patients admitted to the hospital for a stroke within 30 days of being treated and released from the 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"). The measure accounts for the epidemiologic base rate of stroke in the population under study using a risk difference approach (observed [short-term rate] minus expected [long-term rate]).

This measure is not risk-adjusted. The rationale for no risk adjustment is that the developer uses a statistical risk approach (observed short-term stroke risk within 30 days minus expected). The short-term expected rate is estimated in the same patients using the 30-day rate of stroke admission during the period of 91-360 days post-discharge. This attempts to quantify the excess short-term risk of stroke (i.e., attributable risk) due to misdiagnosis. This is not risk-adjusted because, according to the developer, it captures all hospital and patient characteristics and social risk factors.

- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Facility

Reliability

- **SMP ratings for reliability:** H-0; M-5; L-1; I-;2; Measure passes with **moderate** rating

- Reliability testing was conducted at the **measure** score level:
    - A signal-to-noise analysis was conducted.
    - There were concerns by the SMP that this sampling strategy of restricting to hospitals with over 250 cases does not match with how the measure is intended to be implemented.
    - The median reliability score for the entire 967-hospital sample was 0.590 with an interquartile range of 0.414-0.951. Reliability was described by SMP as only "okay", as it was below the 0.7 threshold. It was also marginal/low in hospitals with 250-499 benign dizziness ED discharges (0.582). Results show moderate reliability for hospitals in hospitals that had over 250 dizziness discharges over the three-year time frame. The measure was not reliable below this level of cases. Some SMP comments stated it would not be reliable unless over 500 cases.
    - There was concern regarding a large number of facilities with a reliability score of 1.0; it is unclear whether they had no negative events.
    - There were concerns that for the denominator definition, a single patient may have multiple "index" visits over a three-year period and that this was not accounted for in the reliability testing.
    - There was a concern that only 967 out of 5,503 facilities were included, which only includes EDs with volumes of over 40,000 discharges per year to allow for the 250 threshold benign dizziness cases over the three-year period.

Validity

- **Ratings for validity:** H-0; M-5; L-2; I-1; Measure passes with **moderate** rating
- Validity testing was conducted at the data element level:
    - Data element validity was assessed for two reasons: (1) to test whether stroke diagnoses were valid and (2) to test whether claims were intended to be coded as "benign dizziness" by the clinician when they were coded as such.
    - 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. In this approach, there was a sensitivity for stroke of 86 percent, specificity of 95 percent, PPV of 90 percent, with a kappa agreement of 0.82. In a systematic review of 77 studies, the sensitivity for any cerebrovascular disease was greater than 82 percent in most studies and both specificity and NPV were greater than 95 percent.
    - For denominator reliability for benign dizziness diagnoses, they 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 complaints). Second, they calculated an NPV specifically if another diagnosis was coded; how often did they 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.
    - The observed rate of stroke in 30 days among cases was compared to an "expected rate" to calculate the measure, the latter being 91-360 days after the visit.
    - There were concerns that the expected rate was based on the assumption that the risk of stroke in 91-360 days is not associated with a misdiagnosis of benign dizziness. The SMP also raised concerns that this approach to calculate the "expected rate" fully accounts for risk factors of patients.
    - The lack of risk adjustment for social risk factors was only mentioned in the context of "risk-adjusting away" worse care for racial minorities, and no discussion of potential conceptual relationships.
    - Only a limited sample of hospitals (four hospitals within Johns Hopkins) were used for testing, which may not generalize to other hospitals.

- Only a very small number of hospitals are extremely poor performers, eight out of 927, suggesting that this is a rare event.

Questions for the Standing Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *Do you have any concerns regarding the distribution of reliability scores, particularly when stratified by case volume?*
- *The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Standing Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Standing Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *Do you agree with the developer's approach to social risk factor adjustment?*
- *The Scientific Methods Panel is satisfied with the validity testing for the measure. Does the Standing Committee think there is a need to discuss and/or revote on validity?*

Preliminary rating for reliability:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**
Preliminary rating for validity:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**

Committee Pre-evaluation Comments:

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- Defer to SMP
- some concern
- Problem of the multiple episodes of vertigo before the stroke
- I am not clear about the ability to accurately capture discharge diagnoses that would reflect the target population. Often these are coded as nausea more than dizzy.
- adequate

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

- Defer to SMP
- yes
- yes as above
- Yes
- No

2b1. Validity -Testing: Do you have any concerns with the testing results?

- Defer to SMP
- some concern
- No
- Samples for testing coming from just 4 Hopkins hospitals may not reflect the final cohort
- Na

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- Defer to SMP
- unclear
- issue of patients going to different hospitals for their supposedly benign vertigo and then for their stroke
- Posterior stroke occurs in populations with the traditional risk factors for AIS in general, as well as in younger patients secondary to dissections, etc.
- no risk adjustment

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- The testing hospitals are not representative of the majority of hospitals to which this measure would be applied.
- yes
- identifies meaningful differences in quality
- I need to better understand the completeness of the data
- In addition to benign dizziness as the primary diagnosis code, the developers could have excluded cases where secondary/tertiary sx suggested cause other than stroke

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

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.

- Data elements are available in electronic claims data
- Data collection presents no additional administrative burden

Questions for the Standing Committee:

- *Does the Standing Committee have any concerns related to measure feasibility?*

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

Committee Pre-evaluation Comments:

Criteria 3: Feasibility

3. Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?

- No concerns
- yes
- no concern
- I need to better understand how the target population will be captured given the heterogeneity of presenting symptoms and diagnoses.
- uncertain about uniform use of diagnosis 'benign dizziness.' Provider, facility specific?

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

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

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

4a.1. 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 not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☐ **Yes** ☒ **No**

This is a newly developed measure; therefore, it is not currently being publicly reported or used in an existing accountability program, but developer demonstrates a plan for use for surveillance, public reporting, and program payment purposes. Developer proposes the possible use of an adapted version of the measure that could be used by hospitals for their own internal quality improvement (QI) efforts and also demonstrates various ways in which the measure is currently being used in specific QI programs.

The measure is being implemented at Johns Hopkins as a diagnostic outcome metric in the 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.

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).

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being 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; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others:

Developer states that feedback on the measure from ED physicians in the quality improvement space has been very positive, overall. Additionally, developer states that feedback on the need for balancing measures has been clear and that 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.

Additional Feedback:

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*:*

- *Does the Standing Committee agree that the performance results have been used to further the goal of high-quality, efficient healthcare?*

- *Is there anything that the Standing Committee wishes to discuss related to the current use of the measure?*

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

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results [Impact/trends over time/improvement]

4b2. Benefits vs. harms. 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).

- Developer reports no unexpected findings (positive or negative) during the relatively recent and small-scale deployment of this measure, including no unintended impacts on patients.

Preliminary rating for Usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided?4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- I believe the data sites for ED physician approval is for the intervention that would not be widely available.
- yes
- No concern
- It is not clear to me how this measure has been vetted on a scale and scope reflective of the roughly 5000 Emergency Departments where this would be implemented.
- not currently publicly reported; new measure

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations?4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- I think there are many unintended possible harms here. A large number of people could be admitted and MRIs obtained in an effort to r/o stroke. It would not be necessary for most of them. This VOG requires reading by an oto-neurologist apparently which will not be readily available in most hospitals. It also appears that results are used without even a physical exam.
- yes
- No concern
- This will drive the use of more advanced imaging (MRI) acutely and as an outpatient.
- Unclear, as difficult/impossible to verify if previous discharge was misdiagnosis or unrelated to subsequent stroke

Criterion 5: Related and Competing Measures

Related or competing measures

None

Harmonization

N/A

Committee Pre-evaluation Comments Criterion 5:
Related and Competing Measures

5. Related and Competing: Are **there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?**

- No concerns
- none
- None to my knowledge
- None that I am aware of
- None

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 6/10/2021

- The American Medical Association (AMA) appreciates the opportunity to comment on this measure. While addressing diagnostic error is absolutely critical to ensuring that patients receive the highest quality of care possible, we are concerned with the lack of exclusions such as those patients who leave against medical advice and question whether the measure should be risk adjusted for clinical and/or social risk factors. Specifically, it remains unclear to us whether there are any factors that could contribute to an individual being treated for benign dizziness but then present with an unrelated stroke within the 30 day time window and if this scenario is possible, we believe that the measure should include risk adjustment.

  In addition, we are disappointed to see the minimum measure score reliability results appeared to less than 0.2 according to the histogram included in the testing form. While the median reliability score was 0.590, we believe that measures must meet minimum acceptable thresholds of 0.7 for reliability and the developer should include a minimum case number as a part of the measure specifications to achieve this threshold across all reporting hospitals.

  Lastly, we question whether the information provided as a result of this measure is truly useful for accountability purposes and for informing patients on the quality of care provided by hospitals. Specifically, our concern relates to the relatively limited amount of variation across facilities. While 627 hospitals out of the 967 facilities were identified as performing better than the national average, 0 hospitals performed worse and only 8 were identified as having statistically significant harm. Endorsing a measure that currently only identifies such a small number of outliers does not enable users to distinguish meaning differences in performance and limits a measure's usability.

  We request that the Standing Committee evaluate whether the measure adequately meets the measure evaluation criteria.

- The Federation of American Hospitals (FAH) appreciates the opportunity to comment on this measure. While FAH supports the measure's focus of driving improvements in diagnostic accuracy, we are concerned that the measure may require additional exclusions and question if case minimums to ensure adequate reliability and risk adjustment are needed and whether the measure scores produce sufficient variation to make the results meaningful for accountability purposes.

  The FAH asks the Standing Committee to consider whether some exclusions, delineation of a case minimum, and possible risk adjustment would be appropriate for inclusion in this measure. For example, is it appropriate to hold a facility accountable for a possible missed diagnosis when an individual leaves against medical advice (AMA)? We are also concerned that a minimum number of

patients will be required to ensure that the measure produces acceptable reliability thresholds of 0.7 or higher, yet we were unable to identify any such requirement. Finally, while we appreciate the analyses completed to justify the lack of risk adjustment, we request that the committee discuss whether there are any clinical or social risk factors that could contribute to an individual presenting with a stroke within the 30-day window that is unrelated to the chief complaint of dizziness during the emergency department visit and as a result if there should be some adjustment based on those variables.

The FAH also questions the usefulness of this measure given the limited variation in performance scores with no hospitals identified as statistically worse than the national average, only 8 were identified as having significant harm and the vast majority of the hospitals were no different or better than the national average. We do not believe that this measure provides any new information that would be useful to hospitals and patients.

The FAH asks that the committee carefully consider these concerns during their review.

- **Of the 2 NQF members who have submitted a support/non-support choice:**
  - 1 supports the measure
  - 0 do not support the measure

Combined Scientific Methods Panel (SMP) Scientific Acceptability Evaluation

Evaluating Scientific Acceptability: Instructions

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating. Relevant measure documents are at the bottom of the SharePoint site.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form **if your measure is a composite**.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***INSTRUCTION BOXES*** in comment bubbles to help you answer them.
- ***Please refer to the 2017*** Measure Evaluation Criteria and Guidance document ***(pages 18-24) and the 2-page*** Key Points document ***when evaluating your measures***.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff as soon as possible (methodspanel@qualityforum.org). Is it possible that we can obtain the needed information, but only if requested in a timely manner.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures. Please review the box below to guide your rating.
- If a measure you are evaluating includes multiple measures (e.g., the Hopsital CAHPS measure submsission acutally includes 11 performance measures), all included measures must be rated. You may decide that

one rating applies to all included measures, or you may need to provide separate ratings (e.g., if results are substantially better for one measure than for another).

Measure type	Requirements for Reliability testing	Requirements for Validity testing
Instrument-based measures	BOTH data element and score-level testing	BOTH data element and score-level testing
Composite measures	Score-level testing of the composite measure score; testing of the components is not sufficient	Score-level testing of the composite measure score is desired. At initial endorsement only, empirical or face validity testing of the components OR face validity of the composite is acceptable.
eCQMs	All eCQMs must be tested using the Health Quality Measure Format (HQMF) specifications, which should also use the QDM and value sets published through VSAC Reliance on data from structured data fields is expected; otherwise, unstructured data must be shown to be both reliable and valid. Thus, testing for elements that are not included in structured data fields should be tested at the data element level.	All eCQMs must be tested using the Health Quality Measure Format (HQMF) specifications, which should also use the QDM and value sets published through VSAC Reliance on data from structured data fields is expected; otherwise, unstructured data must be shown to be both reliable and valid. Thus, testing for elements that are not included in structured data fields should be tested at the data element level. Empirical testing is expected, and as of August 2019, data element validation will be required unless justification is provided/accepted. Face validity alone will not be sufficient. Use of a simulated data set (e.g. BONNIE) is no longer accepted for testing validity of data elements
Cost and Resource Use Cost and Resource Use Measure Evaluation Criteria	EITHER data element or score-level testing	Validity is considered in the context of measure intent and threats to validity based on these cost measure-specific components: • Attribution approach • Cost categories • Approach to outliers • Impact of Carve Outs EITHER data element or score-level testing; face validity not accepted for maintenance measures unless justification provided/accepted
All others (Process; Appropriate Use; Structure; Efficiency;	EITHER data element or score-level testing	EITHER data element or score-level testing; face validity not accepted for maintenance measures unless

Measure type	Requirements for Reliability testing	Requirements for Validity testing
Outcome; Intermediate Clinical Outcome; Access)		justification provided/accepted; if data element validity is demonstrated, additional reliability testing is not required

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3614

Measure Title: Hospitalization After Release with Missed Dizzy Stroke (H.A.R.M Dizzy-Stroke)

Measure is:

☒ **New** ☐ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☒ **No**

Submission document: "MIF_3614" document, items S.1-S.22

NOTE*: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

2. **Briefly summarize any concerns about the measure specifications.**

Panel Member 1: For the denominator definition, it is just odd to have multiple "index" visits in the 3-year performance period considered. What happens if the patient has more than one ED visits within the first 360 days – I am assuming that the measure will use the first one in this case, but I did not see that it was explicitly stated. It is somewhat confusing to figure out which is actually the measured entity: is it the facility in which the patient had the first had the index ED visit, or the other facility in which s/he was admitted with stroke diagnosis? From what I can see from the measure description, it is the former. Now, if the patient first goes to hospital A, get discharged from its ED (thus contributing to the denominator). On the 10th day following that ED visit, the patient was admitted to the hospital B with a stroke diagnosis. This measure's intent is to penalize hospital A for potentially missing the stroke diagnosis in the first place. However, I am not sure if there is solid data to prove that hospital A indeed misdiagnosed the patient in the first instance. In particular, claims data on which this measure is based may not have clinical details needed to make such a conclusion. (MIF: S.14) I am thoroughly confused regarding how and why crude delayed 30-day rate per 10,000 visits are calculated, and by corollary how the "attributable" rate (measure) calculation. Particularly, I don't understand, why we have care about things that happen way into the future (90 days through 360 days following index diagnosis). (Testing document 1.5) So, it seems only 967 facilities out of 5,503 facilities were eventually included for the measure, which had at least 250 benign dizziness discharges from ED during 3-yaer period. And these hospitals typically have 40-50K ED discharges a year. This begs the question as to what would is the population of hospital that the measure is attempting to improve quality of? Only larger hospitals with ED discharges >40K per year?

Panel Member 2: None

Panel Member 4: No concerns

Panel Member 5: None

Panel Member 6: I wasn't clear on why the measure is limited to 4 ED events for a single patient - maybe that has something to do with their development data. Otherwise, it makes sense.

Panel Member 8: The index visits require a separation of 360 days for each patient. It seems like that could be 30 days because the numerator is an admission within 30 days of the index. The definition should be consistent with other readmission rate definitions - not a 'show stopper', but interesting choice by the developer.

RELIABILITY: TESTING

Type of measure:

☒ **Outcome (including PRO-PM)** ☒ **Intermediate Clinical Outcome** ☐ **Process**

☐ **Structure** ☐ **Composite** ☐ **Cost/Resource Use** ☐ **Efficiency**

Data Source:

☐ **Abstracted from Paper Records** ☒ **Claims** ☐ **Registry** ☐ **Abstracted from Electronic Health Record (EHR)** ☐ **eMeasure (HQMF) implemented in EHRs** ☐ **Instrument-Based Data** ☐ **Enrollment Data** ☐ **Other (please specify)**

Level of Analysis:

☐ **Individual Clinician** ☐ **Group/Practice** ☒ **Hospital/Facility/Agency** ☐ **Health Plan**

☐ **Population: Regional, State, Community, County or City** ☐ **Accountable Care Organization**

☐ **Integrated Delivery System** ☐ **Other (please specify)**

Measure is:

☒ **New** ☐ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

Submission document: "MIF_3614" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level** ☒ **Measure score** ☒ **Data element** ☐ **Neither**

4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**

5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of **patient-level data** conducted?

☒ **Yes** ☐ **No**

6. **Assess the method(s) used for reliability testing**

Submission document: Testing attachment, section 2a2.2

Panel Member 1: Signal-to-noise analysis

Panel Member 2: Data element: chart review in Hopkins system of accuracy of coding of benign dizziness in ED and stroke on hospital admission. Measure score: S/N

Panel Member 4: The method, a STN model, is appropriate to determine facility level reliability.

Panel Member 5: Performance measure score reliability was calculated using signal-to-noise analysis

Panel Member 6: Time period selected for analysis requires a cross over from ICD-9 to ICD-10. This may be an issue with both reliability and validity testing. Developer states using Adams tutorial for reliability testing, but does not state the actual measurement method used.

Panel Member 8: Signal-to-noise ratio - would help to know the specific statistic or model that they calculated. For example, did they use the beta-binomial model recommended by Adams in the paper they referenced? A little more information would be helpful.

Panel Member 9: Signal to noise analysis was conducted.

7. **Assess the results of reliability testing**

Submission document: Testing attachment, section 2a2.3

Panel Member 1: The median reliability score for the entire 967 hospital sample was 0.590, with an interquartile range of 0.414-0.951.

Panel Member 2: Data element: okay Score: Median S/N 0.59, IQ: 0.414-0.951. Marginal/low reliability in hosps with 250-499 benign dizziness discharges from ED (median 0.582)

Panel Member 4: The sample was limited to facilities who had at least 250 discharges in the denominator. However, this does not match the measure specification and therefore the sample doesn't reflect how the measure will be implemented. The results indicate marginally acceptable results for an outcome measure (median reliability estimate=0.59). There is a curiously high number of facilities with a reliability estimate of 1.0, and it would be helpful to know more about those facilities. As expected, more denominator cases resulted in better reliability, but it is not known how many facilities are in each bucket of discharges shown on page 7.

Panel Member 5: The median reliability score for the entire 967 hospital sample was 0.590, with an interquartile range of 0.414-0.951.

Panel Member 6: Assuming that the reliability measure was appropriate, the results suggest that the reliability is only at an acceptable level for providers with more than 500 index discharges.

Panel Member 8: They report a median reliability statistic of 0.59, which is towards the lower end of our emerging consensus standard. However, I'm more curious about the large number of facilities with a perfect score of 1 - are these situations with no negative events? Also, we see that reliability is very sensitive to number of discharges at a facility -- it looks like the measure is much more reliable (0.81) in facilities with 750 or more dizzy out discharges.

Panel Member 9: The median reliability score for sample of 967 hospitals was 0.590, with IQR of 0.414-0.951. A very large proportion of hospitals had a reliability score of 1.0 which is curious. Based on stratified sample by # of episodes, the median reliability scores were 0.582 for 250-499 “, 0.710 for 500-749, and

0.807 for 750 plus dizzy out" discharges in 3 year performance window. It is unclear given large number of discharges why a 3 year period is needed.

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

Submission document: Testing attachment, section 2a2.2

☒ **Yes**

☒ **No**

☐ **Not applicable** (score-level testing was not performed)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

Submission document: Testing attachment, section 2a2.2

☒ **Yes**

☐ **No**

☒ **Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and **all** testing results):

☐ **High** (NOTE: Can be HIGH **only if** score-level testing has been conducted)

☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has **not** been conducted)

☒ **Low** (NOTE: Should rate **LOW** if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

☒ **Insufficient** (NOTE: Should rate **INSUFFICIENT** if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

Panel Member 1: While the reliability score itself is in the moderate range, I have concerns about whether this measure can reliably assess quality that it intends to measure (See # 2 for my concerns).

Panel Member 2: Might restrict sample to >=500 for adequate reliability, which will limit measure to larger EDs.

Panel Member 4: The developer should, at least, either match up the testing with the specification (in terms of # of discharges) or include reliability results without the restriction of >=250 discharges.

Panel Member 5: Signal to noise was a reasonable approach

Panel Member 6: Rating insufficient because the developer does not state the reliability statistic used to generate the table in 2a2.3.

Panel Member 8: It's a bit hard to characterize the results without more information on the statistic. Since the measure does not perform well for hospitals with fewer dizzy out discharges, I would say this measures is low-to-moderate.

Panel Member 9: Results show moderate reliability for hospitals with more than 250 dizzy out discharges over 3 year timeframe. The measure would not be reliable below this level of cases.

VALIDITY: TESTING

12. **Validity testing level: ☒ Measure score ☒ Data element ☐ Both**

13. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?** *NOTE that data element validation from the literature is acceptable.*

Submission document: *Testing attachment, section 2b1.*

☒ **Yes**

☒ **No**

☒ **Not applicable** (data element testing was not performed)

14. **Method of establishing validity of the measure score:**

☐ **Face validity**

☒ **Empirical validity testing of the measure score**

☒ **N/A (score-level testing not conducted)**

15. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

Submission document: Testing attachment, section 2b1.

☒ **Yes**

☒ **No**

☒ **Not applicable** (score-level testing was not performed)

16. **Assess the method(s) for establishing validity**

Submission document: Testing attachment, section 2b2.2

Panel Member 1: Data element validity through previously published literature that used administrative claims data to identify stroke discharges from acute care hospitals in the U.S by comparing discharge codes against chart abstraction as the gold standard.

Panel Member 2: Measure is rate of admission for stroke in 30 days - rate of admission for stroke in 90-360 days after, on assumption that stroke admissions 90-360 are not associated with misdiagnosed benign dizziness but baseline rate of unrelated stroke in ED patients at that hospital (internal control, rather than risk adjustment)

Panel Member 4: The developer looked at PPV and NPV comparing claims with EHR data using the EHR as the gold standard for about 130 patients. However, I expected the developer to use kappa statistic to understand the chance corrected agreement rates between the data elements.

Panel Member 5: Reasonable approach. Key results from the articles mentioned above are as follows: In the Tirschwell study (PMID: 12364739), the following was found. For ischemic stroke, the sensitivity was 86% (95% CI; 73–94), specificity 95% (95% CI; 88–98), and positive predictive value 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 93% (95% CI 86–97), and positive predictive value 80% (95% CI 64–91), with a kappa score of 0.74. For subarachnoid hemorrhage, the sensitivity was 98% (95% CI 90–100), specificity 92% (95% CI 84–96), and positive predictive value was 86% (95% CI 75–94) with a kappa score of 0.87. The McCormick systematic review (PMID: 26292280) included 77 published 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 stroke 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). In the Kokotailo and Hill study (PMID: 16020772) they found that stroke coding was equally good with ICD-9 (90% correct [95% CI 86-93]) and ICD-10 [92% correct (95% CI 88-95). There were some differences in coding by stroke type, notably with transient ischemic attack, but these differences were not statistically significant.

Panel Member 6: The measure score was not validity tested. The data element testing consists of literature using primarily icd-9 codes and limited to the coding of stroke not the benign dizziness using in the metric definition. Developer also still relying on the 4 hospital set in the previous submission - the data was updated, but still includes only the 4 JHHS hospitals.

Panel Member 8: Excellent approach - I really like the use of EHR data and chart reviews to confirm the diagnostic detail of ED events.

Panel Member 9: For numerator reliability, developers cited 3 studies (2002, 2015 systematic review) and 2005 Canadian study) that evaluated validity of ICD codes for identifying stroke in administrative data. For denominator reliability, they conducted 2 studies focused on code-level validity. Question #1 Positive Predictive Value (PPV): If an ED patient is coded with a "benign dizziness" discharge code, how often do charts suggest the ED provider INTENDED to code a "benign dizziness" discharge? They used Johns Hopkins hospital EPIC HER, complaints, and ED chart notes data from 2 academic hospitals and 2 community hospitals. They calculated the PPV = true positives/all positives for 3 cohorts (dizziness chief complain, oto-vestibular chief complaint, other chief complaint) and reviewed a random sample of 64 charts for the last 2 groups for 2 different performance periods reflecting ICD9 and ICD10 use. Question #2 Negative Predictive Value (NPV): If an ED patient is coded with something OTHER than a "benign dizziness" discharge code, how often do charts suggest the ED provider INTENDED to code something OTHER than a "benign dizziness" discharge? The same data used above were stratified into 2 groups

reflecting high and low risk for misclassification. NPV = true negatives/all negatives. They reviewed a random sub-sample of 67 charts for high-risk sub-group to estimate NPV. To evaluate discharge status, they reviewed 25 random ED patient charts from the 4 hospitals that had a "Discharged" status. They also reviewed a high-risk subset of cases from the numerator.

17. **Assess the results(s) for establishing validity**

Submission document: **Testing attachment, section 2b2.3**

Panel Member 1: While I am generally in agreement with the approach taken by measure developer, I am not confident that the claims data will have all the required clinical details to determine misdiagnosis of stroke in the index case (and the index facility). Moreover, both Tirchwell et al, and McCormic et al. on which the developer is basing many of their arguments with regard to data element validity, both seemed to have used ICD-9 codes only. Are these validation study automatical extends to systems with ICD10 codes?

Panel Member 2: While approach is plausible, would like some documentation or discussion of 30 day and 90-360 day windows. Also, reliability depends on accurate measurement of low baseline rate in 90-360 window. Have graphs that go to 26th week, not 50th, and no data on baseline rates and variation across hospitals.

Panel Member 4: Without kappa scores, there are still unanswered questions about the validity of these data elements.

Panel Member 5: Both the Tirschwell and the McCormick studies found the sensitivity, specificity, and positive predictive values of the ICD-9 stroke codes to be very high (85%+). We found a positive predictive value (PPV) of 100% [CI: 99.89%-100.00%] for coding "benign dizziness." 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 when assessing all the highest-risk cases.

Panel Member 8: Very strong evidence that that diagnosis information on claims is accurately reflecting benign dizziness and discharge status.

Panel Member 9: The articles cited found the ICD items to have high sensitivity and specificity (85%+) indicating claims data can accurately identify patients who have primary stroke dx. They found PPV of 99.997% for coding "not benign dizziness" and PPV of 95.52% in high risk subset. Also found high NPV of 99.9%. Audit of discharge status also found 100% accuracy even when assessing the high risk cases.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

18. **Please describe any concerns you have with measure exclusions.**

Submission document: Testing attachment, section 2b2.

Panel Member 1: None

Panel Member 2: No exclusions

Panel Member 4: No exclusions

Panel Member 5: NA

Panel Member 6: Developer states no exclusions, but the definition of the index ED visit does exclude other ED visits within 360 days. The impact of that exclusion was not assessed.

Panel Member 8: No exclusions, but I'm a little concerned about selection bias -- how often does a hospital meet the minimum criteria of 250 cases of benign dizziness? It is basically the same hospitals year-over-year reflecting something about the types of patients they get or is it different facilities each year?

Panel Member 9: No exclusions

19. **Risk Adjustment**

Submission Document: Testing attachment, section 2b3

19a. **Risk-adjustment method** ☒ **None** ☒ **Statistical model** ☐ **Stratification** ☒**Other**

Panel Member 1: Risk difference approach.

Panel Member 2: Use individual hospital rate of admission for stroke among discharged patients 90-360 days post ED discharge as base rate for population using hospital ED

19b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☒ Yes ☒ No ☒ Not applicable

19c. **Social risk adjustment:**

19c.1 Are social risk factors included in risk model? ☒ Yes ☒ No ☒ Not applicable

19c.2 Conceptual rationale for social risk factors included? ☒ Yes ☒ No

19c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☒ No

19d.**Risk adjustment summary:**

19d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

19d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☒ No

19d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☒ No

19d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration)

☒ Yes ☒ No

19d.5.Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☒ No

19e**. Assess the risk-adjustment approach**

Panel Member 1: Please note that the same person can be in the denominator multiple times, and his or her outcomes will be correlated. The risk difference method that the developer implemented does not seem to have accounted for such correlations.

Panel Member 2: I am ok with using hospital as its own control as a risk adjustment strategy but would like more information on the distribution of the 90-360 baseline rates and their correlation with 0-30 rates.

Panel Member 4: The developer cites a desire to not risk adjust away differences for inequalities in care. However, measure stratification could accomplish this goal while addressing the issue.

Panel Member 5: NA

Panel Member 6: Developer states no risk adjustment, but is using a observed - expected rate is used to measure performance. I cannot locate the definition of expected rate. It appears to be a matched cohort or matched time period method - should this be categorized as stratification?

Panel Member 8: Authors make a strong case for their approach, including no adjustment for race/ethnicity.

Panel Member 9: The rationale for not risk adjusting was that developers use a statistical risk difference approach (observed [short-term stroke risk] minus expected [long-term/baseline stroke risk]). They state that controlling for differences in patients characteristics (case mix) is not needed to achieve fair comparisons across entities with this approach. The short term observed rates is the number of stroke hospitalizations per 10,000 discharges in the first 30 days. The short term expected rate is estimated in exact same patients using the average 30-day rate of stroke admission during period 91-360 days post discharge to estimate the base rate of stroke. 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. This presumably captures all of the hospital patient characteristics and social risk factors according to the developers.

20. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

Submission document: Testing attachment, section 2b4.

Panel Member 1: I am concerned that this measure may not measure what it is intending to measure. The 3-year horizon (performance period) is very long, and by construction, its target population is the only large hospitals with ED discharging >40K ED patients yearly. In an ideal world, the most potent way for validating this measure would have been that the measure developer could verify their estimated measure with the actual numbers from the facilities included in the final sample. However, as indicated in the testing form, due to privacy issues, the dataset used for testing does not allow identifying the facilities included in the measure.

Panel Member 2: I did not see data on variation in rates across sample

Panel Member 4: The developer's analysis shows that a very small number of hospitals were extremely poor performers.

Panel Member 5: None. 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 25th percentile is close to the difference between the median and the 75th percentile). With

the measure we were also 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". As described above in 2a2.4, we expect that this same measure, used in clinical practice with a more complete data set reflecting all ED dizziness discharges (rather than only the 20% Medicare fraction), would demonstrate even greater precision to identify differences among facilities.

Panel Member 6: There is significant skew in the scores. 65% of the hospitals are 'better' than the nation after; none worse. This measure does not provide meaningful comparisons to measure performance.

Panel Member 8: Only 8 out of 927 facilities had a high rate of misdiagnosis - I guess that's good if you are experiencing dizziness and go to an ED. However, it does seem like the measure is focused on a relatively rare event.

Panel Member 9: They found significant variation between facilities on the calculated measure, with performance fairly evenly distributed around the median performance (i.e., difference between the median and 25th percentile is close to the difference between the median and the 75th percentile). They were also able to identify a sizable number of facilities who are "better than the national average" (627/967), and were able to identify a small number of facilities (8/967) that had statistically significant rates of misdiagnosis "harm". Interestingly, this approach found 0/967 hospitals identified as "worse than the national average" which raises concerns.

21. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**

Submission document: Testing attachment, section 2b5.

Panel Member 1: None

Panel Member 2: No discussion of variation in coding for benign dizziness across EDs.

Panel Member 4: N/A

Panel Member 5: NA

Panel Member 6: NA

22. **Please describe any concerns you have regarding missing data.**

Submission document: Testing attachment, section 2b6.

Panel Member 1: None

Panel Member 2: NA

Panel Member 4: The developer used a very synthetic/massaged data source so this is not possible to assess with the data used for the evaluation.

Panel Member 5: None

Panel Member 6: Developer did not analyze impact of missing data.

Panel Member 8: None

For cost/resource use measures ONLY:

23. **Are the specifications in alignment with the stated measure intent?**

☒ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

24. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

25. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

☒ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☒ **Low** (NOTE: Should rate LOW if you believe that there **are** threats to validity and/or relevant threats to validity were **not assessed OR** if testing methods/results are not adequate)

☒ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level **is required**; if not conducted, should rate as INSUFFICIENT.)

26. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

Panel Member 1: My rating is based on my assessments of validity subcriteria in #12 through #24.

Panel Member 3: Best practice

Panel Member 4: Lack of kappa statistics prevents analysis of the chance adjusted agreement rates.

Panel Member 5: Limited testing, as new measure, but what was completed is reasonable.

Panel Member 6: See answers to #19e and #20.

Panel Member 8: I was impressed with the EHR and chart abstraction analysis used to validate event characteristics.

Panel Member 9: It is not clear that the "risk difference approach" can sufficiently capture the "expected rate" of stroke, or that subtracting the calculated expected rate fully accounts for risk factors of patients. The lack of social risk adjustment was mentioned only in context of not "risk adjusting away" worse care for racial minorities, but no discussion of potential conceptual relationships (which may or may not exist). Only 4 Johns Hopkins hospitals were used to validate the data elements which may not reflect the universe of hospitals in the US. No hospitals were found to perform worse than national average.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

☐ **High**

☐ **Moderate**

☐ **Low**

☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

Panel Member 2: No data on variation, opportunities for improvement, miscoding of benign dizziness outside of Hopkins system, value of measure.

Developer Submission

NQF #: 3614

Corresponding Measures:

De.2. Measure Title: Hospitalization After Release with Missed Dizzy Stroke (H.A.R.M Dizzy-Stroke)

Co.1.1. Measure Steward: Johns Hopkins Armstrong Institute for Patient Safety and Quality

De.3. Brief Description of Measure: This outcome measure tracks the rate of patients admitted to the hospital for a stroke within 30 days of being treated and released from the 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"). The measure accounts for the epidemiologic base rate of stroke in the population under study using a risk difference approach (observed [short-term rate] minus expected [long-term rate]).

1b.1. Developer Rationale: Diagnostic error is a major public health problem.[1] The lack of operational measures is a critical barrier to improving diagnosis.[2,3] 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.[4] Among vascular events, missed stroke is the leading cause of serious harm to patients. Misdiagnosis of stroke disproportionately occurs when symptoms and signs are not typical or obvious.[5,6] Among strokes, the most commonly missed clinical presentation is patients presenting dizziness or vertigo, easily mistaken for inner ear disease.[5] In US emergency departments (ED) each year, an estimated 45,000-75,000 patients with strokes presenting dizziness or vertigo are missed and erroneously discharged.[6]

ED patients with acute dizziness and vertigo could be diagnosed correctly using evidence-based bedside examinations,[7,8] but there is currently a large evidence-practice gap[9] in ED diagnosis, resulting in substantial harms to patients.[6] Without timely, accurate diagnosis, these patients suffer misdiagnosis-related harms[10] from lack of prompt treatment.[5] The most common harm is preventable major stroke after minor stroke or transient ischemic attack (TIA),[11,12] with major stroke leading to subsequent hospitalization. Crude short-term stroke hospitalization rates per 10,000 dizziness discharges from the ED vary at least from 20-80.[6] Adjusting for baseline stroke risk across groups does not eliminate practice variation.[13]

This outcome measure tracks the rate of patients admitted to the hospital for a stroke within 30 days of being treated and released from the 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"). This measure is the first operationally-viable performance measure of stroke misdiagnosis for the hospital setting. Hospital EDs will be able to use the measure internally to track their performance over time, as they work to implement interventions to reduce misdiagnosis of strokes. The measure can also be used by external entities for public reporting and pay-for-performance, as external pressures to encourage hospital improvements.

S.4. Numerator Statement: The number of ED index visits during the performance period that are followed within 30 days by an inpatient hospital admission to any hospital that ends in a primary diagnosis of stroke.

S.6. Denominator Statement: Patients discharged from the ED with "benign dizziness" as the primary diagnosis code, counting a patient's first such discharge during the performance period (an "index visit) and all subsequent such discharges that fall outside a 360-day follow-up window from the previous qualifying "index visit".

S.8. Denominator Exclusions: The measure has no exclusions. All patients discharged from the ED with "benign dizziness" as their primary diagnosis code are included in the measure denominator.

De.1. Measure Type: Outcome

S.17. Data Source: Claims

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

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

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

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?

1. Evidence and Performance Gap – Importance to Measure and Report

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

3614_NQF_evidence_attachment_FINAL_210402-637529982255843202.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 3614

Measure Title: Hospitalization After Release with Missed Dizzy Stroke (H.A.R.M Dizzy-Stroke)

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: **4/2/2021**

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Outcome: **Stroke misdiagnosis-related harm (diagnostic adverse event)**

☐ Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)

☐ Intermediate clinical outcome (*e.g., lab value*):

☐ Process:

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or 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.

Introduction

An estimated 250,000 hospitalized patients are harmed by diagnostic error each year in the US,[1] including perhaps 80,000 preventable deaths annually[2] and likely a comparable amount of serious disability.[3] The aggregate estimate for serious misdiagnosis-related harms across clinical settings is likely much higher.[4] Physician-reported errors and closed malpractice claims indicate that stroke is among the most frequent causes of serious misdiagnosis-related harms.[5]

It is important to note that strokes are not misdiagnosed when they are obvious (e.g., paralysis on one side of the body and inability to speak) --- they are misdiagnosed when they are **not** obvious (e.g., because they look like something else). A recent systematic review of emergency department (ED) stroke misdiagnosis found the highest odds of misdiagnosis are conferred by a presenting stroke symptom of dizziness or vertigo (OR 14).[6] This is because strokes in the back part of the brain have symptoms and signs that very closely mimic debilitating (yet non-life-threatening) inner ear diseases.

Envisioned below is scenario in which an initial misdiagnosis is identified through a biologically plausible and clinically sensible temporal association between an initial symptomatic visit (that ended with a benign diagnosis rendered) and a subsequent revisit (that ended with a dangerous diagnosis confirmed):

> Scenario Step 1: The patient presents at an Emergency Department (ED) with a chief complaint of dizziness or vertigo. This sparks an ED investigation of possible causes, the majority of which are ultimately deemed benign (particularly inner ear diseases).
>
> Scenario Step 2: The patient is evaluated by the health care team, and a diagnosis of "benign/non-specific dizziness" is rendered, after which the patient is discharged home. Events that occur subsequent to this treat-and-release visit make it clear this diagnosis was in error.
>
> Scenario Step 3: The patient returns to the hospital within a short timeframe (e.g., 30 days) with continued/worsening symptoms and is admitted to the hospital where the diagnosis of stroke is confirmed. At the end of the hospitalization, the patient is discharged with a diagnosis of stroke.

The failure to correctly diagnose the underlying disease(s) in a timely manner may be followed by illness progression that might have been avoided through prompt diagnosis and treatment (potentially preventable misdiagnosis-related harms). Specifically, there is a well-established acute risk profile for major stroke following so-called minor stroke and transient ischemic attacks (TIA) (Figure below). That known major stroke risk profile (highest in the first 72 hours and leveling off by 90 days) happens to match almost exactly the risk profile of stroke hospitalization after ED "benign dizziness" discharge.

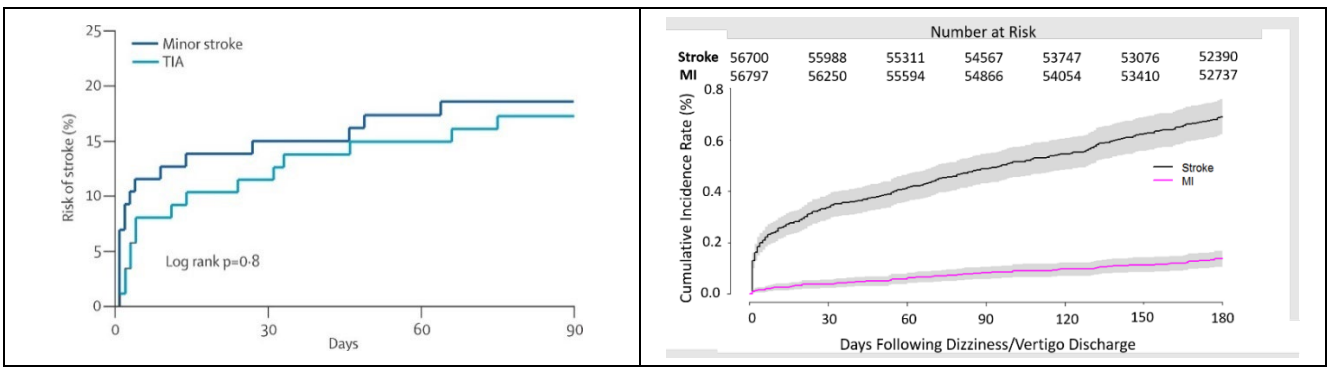


Figure. Cumulative incidence curve for natural history of major stroke following minor stroke or TIA. Data are from the population-based Oxford Vascular Study as represented in Rothwell, et al.[7]	**Figure. Cumulative incidence of stroke hospitalizations post ambulatory (ED or other) treat-and-release as "benign dizziness."** Heart attack (MI) returns are shown as a "control" comparator to demonstrate the specificity of the association. Data are from a collaboration between Johns Hopkins and Kaiser Permanente Mid-Atlantic States.[8]

Here we use the terminology "benign dizziness" to refer to a treat-and-release ED visit diagnosis of inner ear disease or non-specific dizziness. We do **not** use the term "benign" to discount the (often intense) suffering felt by inner ear disease patients, nor to imply that those discharged by ED physicians with a primary, symptom-only diagnosis of "dizziness, not otherwise specified" are presumed to have had a "complete" work-up or "final" diagnosis not necessitating additional outpatient medical follow-up. We **do** use the term "benign dizziness" as shorthand for patients in whom a stroke diagnosis was clinically excluded and/or deemed so unlikely as to not warrant hospital admission.

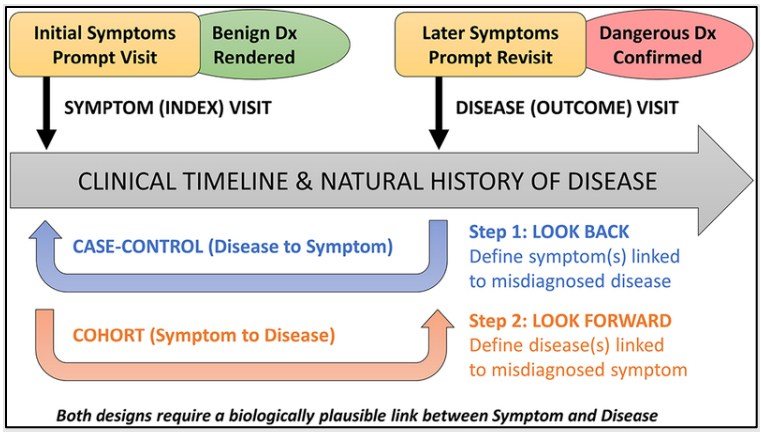
While the severity or preventability of such misdiagnosis-related harms cannot be ascertained merely by the presence of a hospitalization post treat-and-release discharge, there is ample convergent evidence from multiple sources that ED errors in diagnosis with dizziness and vertigo are frequent[6,9,10] and that some patients suffer serious, irrevocable harms from missed opportunities to treat strokes, such as the devastating "locked in syndrome" (quadriplegia plus mutism with intact consciousness).[11,12] Preventable adverse outcomes of misdiagnosis result from missed opportunities for thrombolysis,[13,14] early surgery for malignant posterior fossa edema,[15,16] or prevention of subsequent infarction.[7,17,18] Rapid treatment improves health outcomes[19,20] and prompt prophylaxis lowers repeat stroke risk by up to 80%.[21,22] Thus, patients generally benefit from early, correct diagnosis.

Note that in this application the symptoms "vertigo" (a false sense of motion) and "dizziness" (spatial disorientation without a false sense of motion) are used interchangeably, since most ED patients cannot reliably distinguish between the two symptoms in the throes of an acute attack[23] and because the symptom type is, at best, a very weak predictor of an underlying stroke etiology.[24,25]

SPADE Conceptual Framework

Symptom-Disease Pair Analysis of Diagnostic Error (SPADE) is a conceptual framework and methodological approach for measuring diagnostic quality and safety that is based on the notion of change in diagnosis over time.[26] Given what is known about disease natural history and pathophysiology, the model identifies misdiagnosis-related harms based on time-linked markers of diagnostic delay that are clinically sensible, biologically plausible and specific to symptom-disease pairs.

The framework shown below illustrates differences in structure and goals of the 'look back' (disease to symptoms) and 'look forward' (symptoms to disease) analytical pathways. The 'look back' approach defines the spectrum of high-risk presenting symptoms for which the target disease is likely to be missed or misdiagnosed; the 'look forward' approach defines the frequency of diseases missed or misdiagnosed for a given high-risk symptom presentation.



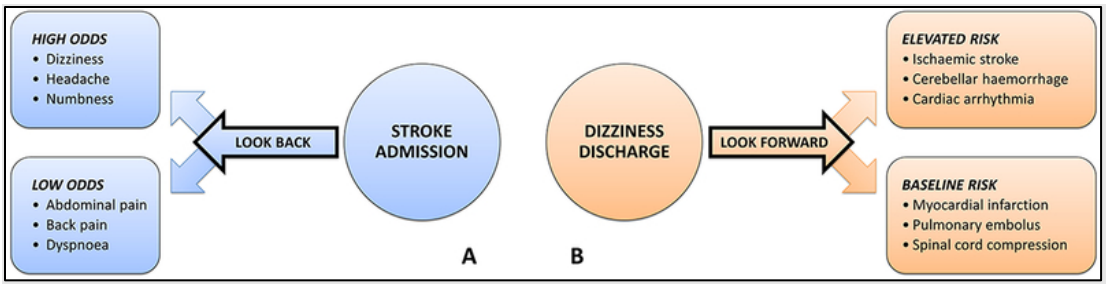
Symptom-disease pairs that may be 'diagnostic error dyads' can be analyzed using either a 'look-back' or a 'look-forward' approach (figure below).[26] The look-back approach takes an important *disease* and identifies which clinical presentations of that disease are most likely to be missed. The look-forward approach takes a common *symptom* and identifies which important diseases are likely to be missed among patients who present with this symptom. When little is known about misdiagnosis of a particular disease, a look-back analysis helps identify promising targets to establish one or more diagnostic error dyads. Once one or more diagnostic error dyads are established, a look-forward analysis can be performed to measure real-world performance. Thus, this conceptual framework makes it possible to identify dyads at high risk for misdiagnosis-related harms. While stroke is one such example, we believe this methodological approach can be applied to any acute disease (e.g., heart attack[27,28] sepsis[29]).

To our knowledge, there are no valid, reliable measures of misdiagnosis-related harms that have yet been approved by the NQF.[30] However, the National Academy of Medicine and NQF have both made clear that such measures are desperately needed in order to promote diagnostic excellence.[30-32] In a recent report, the NQF labeled measures related to dizziness and stroke as one of just three measures of diagnostic safety and quality that are "important and feasible for development now" for ED care.[32] We believe that the dizziness-stroke dyad is ideal for applying the SPADE methodology,, given the strong underlying evidence base and solution sets that have recently become available, such as remote tele-specialty consultation (see 1a.2).

Method for Establishing a Symptom-Disease Pair

Using dizziness-stroke as the example, the 'look-back' approach (A) is used to take a single disease known to cause harm (e.g., stroke) and identify a number of high-risk symptoms that may be missed (e.g., dizziness/vertigo).[26] In this sense, the 'look-back' approach (case-control design) can be thought of as hypothesis generating. In this example, stroke is chosen as the disease outcome. Various symptomatic clinical presentations at earlier visits are examined as exposure risk factors, some of which are found to occur with higher-than-expected odds in the period leading up to the stroke admission.

The 'look-forward' approach (B) is used to take a single symptom known to be misdiagnosed (e.g., dizziness/vertigo) and identify a number of dangerous diseases that may be missed (e.g., stroke). In this sense, the 'look-forward' approach can be thought of as hypothesis testing. In this example, dizziness is chosen as the exposure risk factor, and various diseases are examined as potential outcomes, some of which are found to occur with higher-than-expected risk in the period following the dizziness discharge.



1a.3 Value and Meaningfulness: **IF** this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

N/A

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

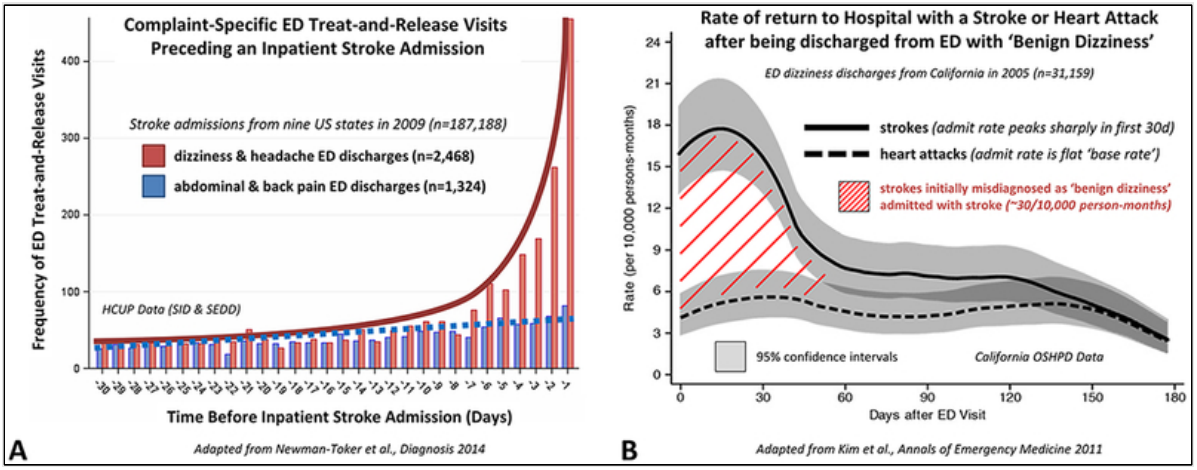
1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

Evidence for Misdiagnosis of Emergency Department Dizziness, Including that Caused by Stroke

There is a large body of literature supporting the notion that dizziness is frequently misdiagnosed in the ED, both with respect to stroke and with respect to inner ear (vestibular) disorders (reviewed in Kerber & Newman-Toker, 2015[10]). Misdiagnosis rates as high as 80% have been documented.[33] This literature includes prior work using SPADE or similar methods to demonstrate empirical, statistically-valid evidence of measurable excess short-term hospitalizations for stroke using both look back[34] and look forward methods.[8,35] These same results for stroke and dizziness have been demonstrated by multiple teams around the world using multiple data sources, including prospective studies at community EDs,[36] integrated health systems,[8,35] state/regional databases,[34,37-39] and national data from Taiwan[40] and the US (Medicare data below).

Patients hospitalized for stroke (n~190 000 admissions from 9 US states in 2009) are more likely to have had a treat-and-release ED visit for so-called 'benign' dizziness within the prior 14 days than have had an ED visit for a different chief complaint (Figure A, adapted from Newman-Toker et al, *Diagnosis,* 2014[34]). Using the 'lookback' approach, dizziness is an over-represented symptom (i.e., among patients with inpatient stroke admissions, the odds of a recent ED discharge with benign dizziness are higher than the average frequency for this symptom in an overall ED population). Treat-and-release ED dizziness discharges occur disproportionately in the days and weeks immediately prior to stroke admission, in a biologically plausible and clinically sensible temporal profile (exponential curve before admission, shown in red) paralleling the natural history of major stroke following minor stroke or transient ischemic attack (TIA). In contrast, abdominal and back pain discharges are under-represented (i.e., among strokes, low odds of a recent ED discharge) and temporally unassociated to the stroke admission (Figure A). This indicates the association seen with dizziness is not simply "nonspecific."

'Benign' dizziness treat-and-release discharges from the ED (n~30 000 visits per year) are more likely to return for an inpatient stroke admission within the subsequent 30 days (Figure B, adapted from Kim et al, Annals of Emergency Medicine, 2011[37]). Using the 'look-forward' approach, stroke turns out to be the disease with the most elevated short-term risk profile (i.e., among patients discharged from the ED with supposedly benign dizziness, the greatest rate of subsequent stroke admission relative to other common symptoms). These occur disproportionately in the days and weeks immediately following the dizziness discharge from the ED, again in a biologically plausible temporal profile ('hump' seen after discharge, shown as red hatched area), paralleling the natural history of major stroke following minor stroke or TIA. By contrast, heart attack risk remains at baseline (i.e., among dizziness discharges, there is a low, stable rate of myocardial infarction admissions over time) and is temporally unassociated to the initial ED dizziness discharge (Adapted from Kim et al., *Annals of Emergency Medicine*, 2011). Likewise, this also indicates that the association with dizziness is not simply "nonspecific."



Note that in panel B the difference in shape of the exponential decay curve (vs. that seen in panel A and other graphs) is likely related to bin size in statistical analysis, rather than a true property of the data. We have varied bin size and seen similar shaped curves when using larger bins.

Not surprisingly, these same patterns are evident in US Medicare data. Using the same data analyzed for the purposes of the proposed measure, but examining over 10 years, we see the following graph.

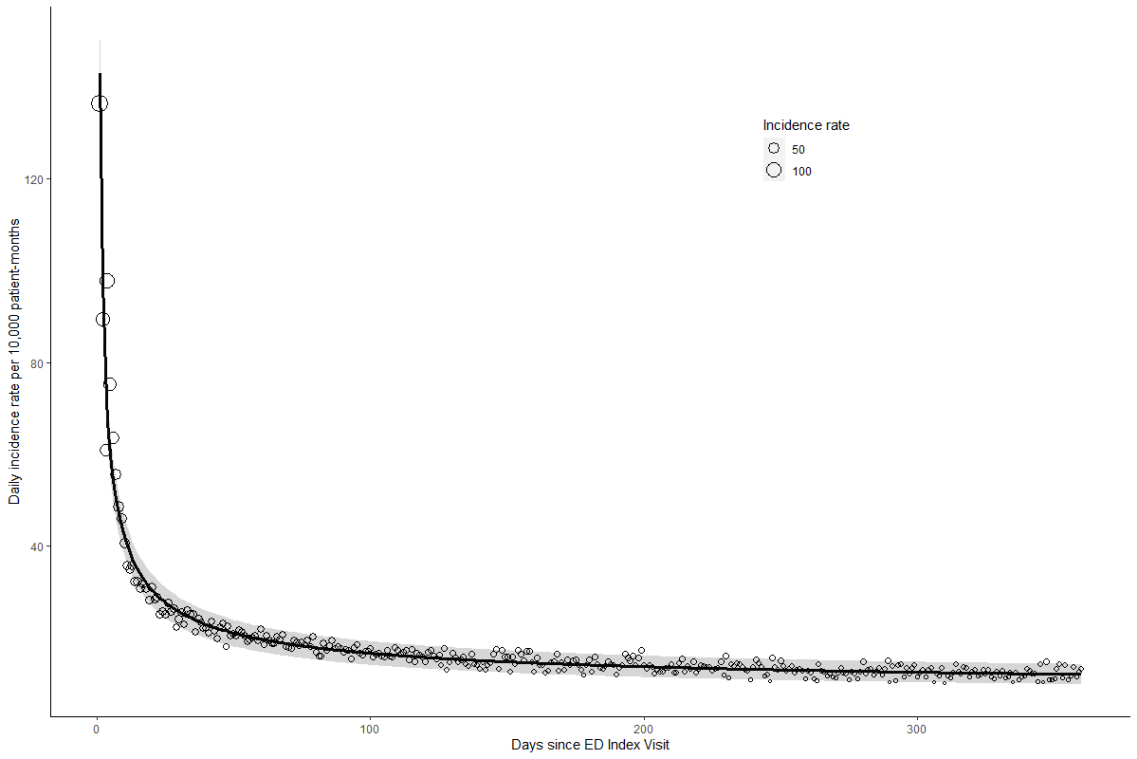


Figure. US Medicare data (10 years from 2009-2018, inclusive) showing the rate of return (incidence rate) for a stroke hospitalization after a "benign dizziness" treat-and-release discharge from the ED. Note that the rate begins high (>100 per 10,000 = >1%) and rapidly declines to a roughly stable baseline rate. Given that there are approximately 1.5M such discharges annually in the US, this represents a substantial number of patients whose cerebrovascular events were misdiagnosed, leading to potential harms including disability or death for tens of thousands of Americans annually through missed opportunities for acute treatment or early secondary prevention. Note that there are probably 5-fold more missed strokes, since only 15-20% of those with TIA or minor stroke will go on to worsening/major stroke requiring hospitalization. *Gray shading reflects the 95% confidence interval.*

Clinical Validity of the Dizziness-Stroke Association Based on Subgroup Analysis of Strokes

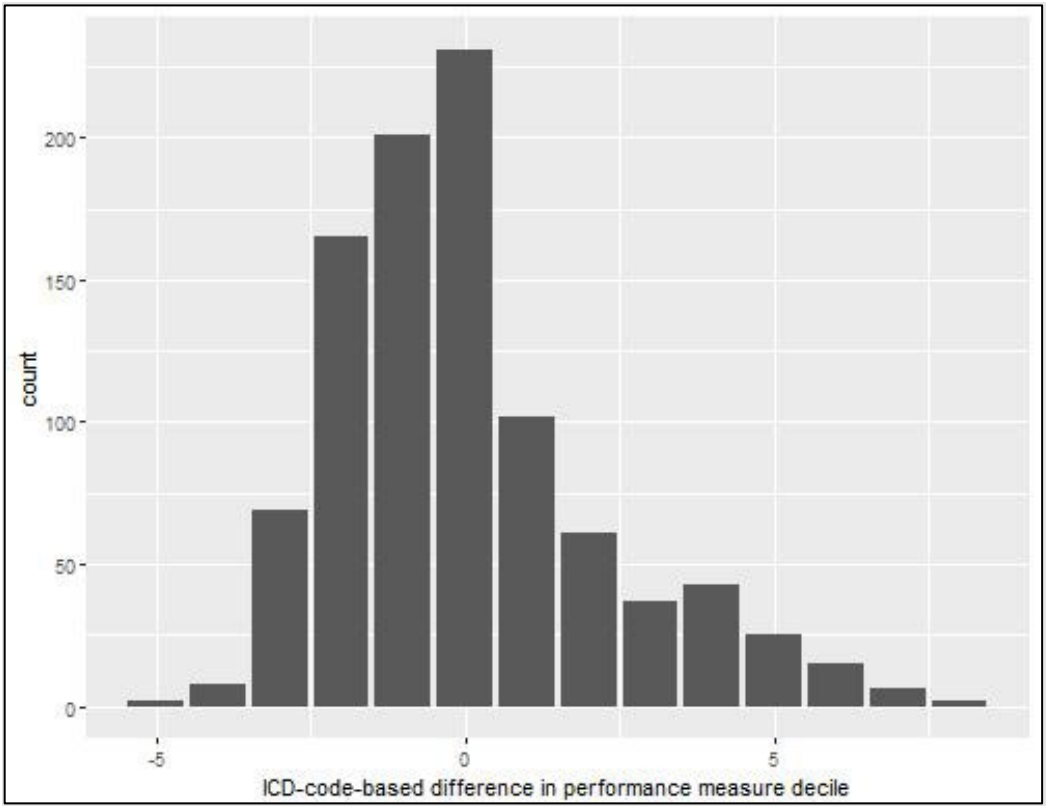
There are multiple types of data that support the validity of the causal relationship between dizziness and missed stroke (reviewed in Kerber & Newman-Toker, 2015[10]). However, even using these large-scale, administrative data, we can further demonstrate the validity of these findings.

First, there is the comparison between all cerebrovascular events and just acute ischemic strokes. Our proposed measure uses a broad panel of cerebrovascular codes which were based on the Elixhauser classification system (HCUP CCS)[41] available at https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (for ICD-9-CM) and https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp (for ICD-10-CM). It is clinically known that cerebellar hemorrhages are only rarely mistaken for benign dizziness,[42] and the majority of missed strokes in dizziness are known be ischemic.[34,43] Thus, focusing the analysis on a narrower subgroup of codes related just to acute ischemic stroke (as defined by Tirschwell et al.[44] [ICD-9-CM] and Hsieh et al.[45] [ICD-10-CM]), which sacrifices some sensitivity for relevant events[44,46] should, in theory, produce a similar result in terms of measure performance to that seen with all strokes.

When using the narrower set of acute ischemic stroke codes below…

- ICD-9-CM code 433.x1 ("x," the fourth digit, can vary to specify a specific arterial distribution), 434 (excluding 434.x0), or 436. [from Tirschwell et al., 2002]
- ICD-10-CM code I63 (any ending digits) [from Hsieh et al., 2020]

… for the three-year performance window from 2015-01-01 to 2017-12-31, we found that ~77.5% of pairwise rankings are reserved comparing two sets of ICD codes (i.e., broader HCUP vs. narrower Tirschwell/Hsieh codes) (Kendall's tau is 0.55 [p-value < 0.01]). The histogram below reflects the extent of change in decile ranking across the 967 facilities with at least 250 visits over 3 years.



Second, there is the examination of results based on acute ischemic stroke vascular territory. It is well known clinically that dizziness caused by stroke is most often due to ischemia in the posterior (vertebrobasilar) circulation, as opposed to the anterior (carotid) circulation.[43] ICD-10 codes offer an additional window into the measure's validity, since vascular territory can now readily be recorded (though, in current practice, it rarely is). It **should** be the case that stroke hospitalizations after dizziness are disproportionately posterior circulation strokes, as opposed to anterior circulation strokes. It should **not** be the case that strokes are exclusively posterior circulation in location, because some underlying etiologies lead to cerebrovascular events in either/both vascular territories (e.g., cardioembolic strokes from atrial fibrillation) and because anterior circulation events can also cause dizziness or vertigo.[47]

We examined anterior vs. posterior circulation strokes for the entire ICD-10 period (performance window 2015-10-01 to 2017-12-31) for all sites lumped together. Dizziness and stroke visits were chosen as if we were using them to calculate the Avoid H.A.R.M measure (i.e., we required number of enrollment days and count repeated stroke visits only if they are at least 360 days apart). In aggregate, there were 1,546 anterior circulation strokes and 595 posterior circulation strokes, with 288,975 unspecified (or classified as due to an

etiology that could involve either territory). The results in the Table below clearly demonstrate that **posterior circulation strokes are overrepresented in the 30 days following the ED index visit and the Avoid H.A.R.M. measure appropriately reflects this fact.**

Category	Anterior Circulation	Posterior Circulation	% Posterior
Total strokes coded as either anterior or posterior (without regard to dizziness)	1,546	595	27.8%
Strokes <30d post ED dizziness discharge	67	76	53.1%
Strokes 91-360d post ED dizziness discharge	359	106	22.8%
Avoid H.A.R.M. measure per 10,000 ED discharges (95% CI)	1.10 (0.48, 1.72)*	2.10 (1.48-2.71)*	- -

** Note that the absolute values are very low because very few stroke events are coded this way*

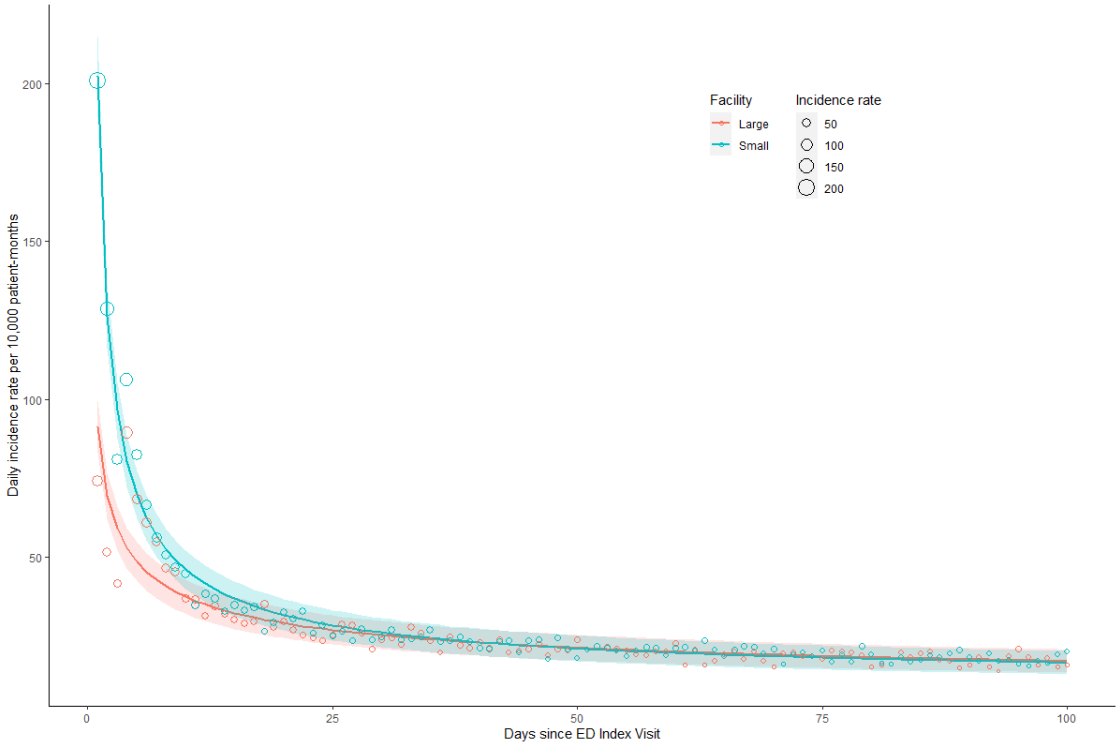
- - cell intentionally left blank

Measure/Outcome Relationship to Structure, Process, or Intervention

There are multiple factors known to predispose to diagnostic error measured in this way, including patient demographic factors (e.g., non-white, female[34]), diagnostic process (e.g., reliance on negative CT neuroimaging to rule out stroke[38]), and fixed or dynamic attributes of healthcare institutions (e.g., non-teaching, low annual ED volume, and high discharge fraction on the day of the encounter[34]).

Although not as strong as some other factors such as the high discharge fraction on the day of the encounter (6.34-fold greater odds),[34] low annual ED volume is associated with 57% greater odds of a missed stroke.[34] We were able to analyze this variable in relation to our measurement approach for this application as a means of further demonstrating a relationship to an institutional structural variable.

To further validate that this known issue of low annual ED volume is a risk factor for missed stroke in dizziness, the graph below shows the stroke incidence rate over 1-100 days after ED index visit, with 95% CIs, for 5472 facilities (10-year window 2009-2018). A cutoff of 1,000 ED index visits over ten years was used to define large vs. small facilities (1,472,612 ED index visits are in large facilities, and 1,422,724 in small facilities). The Figure below illustrates that small facilities have higher short-term strokes. **With more complete data sets than Medicare (e.g., local hospital data, which have 100% of visits, rather than only ~20%), these differences could be measured using the proposed 3-year rolling window.**



Finally, the panel may wonder whether any interventions can influence these events or outcomes. As yet, we do not have clinical trials data demonstrating a statistically-significant reduction in stroke hospitalizations post dizziness discharge (a large trial [10,000 subjects] has been planned for this purpose and is currently under review at PCORI). However, we do have strong evidence that ED diagnosis of dizziness can be dramatically improved through specialty consultation. This evidence comes both from our AVERT phase II clinical trial (Clinical Trials.gov ID NCT02483429) and our clinical "Tele-Dizzy" consultation service.[48] Relevant results from both are shown below in tabular form.

AVERT TRIAL: Final multidisciplinary panel adjudication after follow-up is now complete for 83 of 130 randomized patients from the AVERT trial. This masked adjudication by a dizziness specialist (oto-neurologist), has substantially greater accuracy than the ED providers overall and, especially, for vestibular disorders (Table 3). This diagnosis is performed with access to only results of video-oculography (VOG) quantitative eye movement recordings and a brief patient history from the index ED visit (e.g., 56yo woman with new dizziness for 12 hours, vomiting, difficulty walking).

Table 3. AVERT diagnostic accuracy by ED clinical (all tests) vs. specialist using only eye movements by VOG

Final Diagnosis after Follow-up by Multidisciplinary Panel	ED Clinical Diagnostic Accuracy, % (95% CI)	Oto-neurologist Using VOG-only, % (95% CI)	Relative Change in Diagnostic Accuracy	p-value (exact or Chi-square test)
Most Common Inner Ear (n=41)				
BPPV (n=24)	20.8% (7.1-42.2)	70.8% (48.9-87.4)	↑ 240%	<0.0001
Vestibular neuritis (n=17)	35.3% (14.2-61.7)	76.5% (50.1-93.2)	↑ 117%	0.0004
All Causes (n=83)				
3-Category* Schema (n=83)	38.6% (28.1-49.9)	68.8% (57.4-78.7)	↑ 78%	<0.01
6-Category* Schema (n=83)	31.3% (21.6-42.4)	52.5% (41.0-63.8)	↑ 68%	0.01
12-Category* Schema (n=83)	28.9% (19.5-39.9)	51.2% (39.8-62.6)	↑ 77%	0.01

** Number of diagnostic categories (3 = central, peripheral, other; 6 = posterior canal BPPV, vestibular neuritis, other peripheral, central, other, unknown; 12 = 7 specific peripheral vestibular categories, 3 central categories, medical/psychiatric, unknown)*

TELE-DIZZY CONSULTATION SERVICE: Specialists evaluate ED dizzy patients remotely with the aid of VOG recordings. Results from the initial implementation of this program are in Table 4. Provider satisfaction with

the program is extremely high (93% of ED physicians are "very satisfied"). Most of the minor ED complaints stem from the fact that the service is not available on evenings or weekends. The Tele-Dizzy service has not yet experienced a single missed stroke (i.e., hospitalization within 30 days).

Table 4. Results of Tele-Dizzy pilot program at Johns Hopkins Hospital (n=287 teleconsults)

Category	Parameter	Baseline*	Tele-Dizzy	Improvement	p-value (χ^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 our co-primary stroke outcome measure would be in VERTIGO (not using the comparator population for Tele-Dizzy). The Tele-Dizzy value is based on actual patients seen at the same hospital – thus far, there have been zero stroke returns.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of 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. (IOM)

☐ 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

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	*
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If	*

Systematic Review	Evidence
not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	*
Provide all other grades and definitions from the evidence grading system	*
Grade assigned to the **recommendation** with definition of the grade	*
Provide all other grades and definitions from the recommendation grading system	*
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	*
Estimates of benefit and consistency across studies	*
What harms were identified?	*
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	*

*cell intentionally left blank

_____

1a.4 OTHER SOURCE OF EVIDENCE

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

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1. Gunderson CG, Bilan VP, Holleck JL, Nickerson P, Cherry BM, Chui P, Bastian LA, Grimshaw AA, Rodwin BA. Prevalence of harmful diagnostic errors in hospitalised adults: a systematic review and meta-analysis. BMJ Qual Saf. 2020;29(12):1008-18.

2. Leape LL, Berwick DM, Bates DW. Counting deaths due to medical errors (in reply). JAMA. 2002;288(19):2404-5.

3. Saber Tehrani AS, Lee H, Mathews SC, Shore A, Makary MA, Pronovost PJ, Newman-Toker DE. 25-

Year summary of US malpractice claims for diagnostic errors 1986-2010: an analysis from the National Practitioner Data Bank. BMJ Qual Saf. 2013;22(8):672-80.

4. Newman-Toker DE, Nassery N, Schaffer AC, Yu-Moe CW, Saber Tehrani AS, Clemens GD, Wang Z, Zhu Y, Fanai M, Siegal D. Burden of Serious Misdiagnosis-Related Harms in the United States—Population-Based Estimate Using the "Big Three" (Vascular Events, Infections, and Cancers). Diagnostic Error in Medicine, 11th Annual Conference; New Orleans, LA2018.

5. Newman-Toker DE, Schaffer AC, Yu-Moe CW, Nassery N, Saber Tehrani AS, Clemens GD, Wang Z, Zhu Y, Fanai M, Siegal D. Serious misdiagnosis-related harms in malpractice claims: The "Big Three" - vascular events, infections, and cancers. Diagnosis (Berl). 2019;6(3):227-40.

6. 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):1468-77.

7. Rothwell PM, Buchan A, Johnston SC. Recent advances in management of transient ischaemic attacks and minor ischaemic strokes. Lancet Neurol. 2006;5(4):323-31.

8. Nassery N, Mane K, Liu F, Sangha NX, Sharp AL, Shamim EA, Rubenstein KB, Kharrazi H, Nagy P, Kanter MH, Horberg M, Ford D, Pronovost PJ, Newman-Toker DE. Measuring missed strokes using administrative and claims data: towards a diagnostic performance dashboard to monitor diagnostic errors. Diagnostic Error in Medicine 2016; November 6-8, 2016; Hollywood, CA.

9. Newman-Toker DE. Missed stroke in acute vertigo and dizziness: It is time for action, not debate. Ann Neurol. 2016;79(1):27-31.

10. Kerber KA, Newman-Toker DE. Misdiagnosing dizzy patients: Common pitfalls in clinical practice. Neurol Clin. 2015;33(3):565-75.

11. Savitz SI, Caplan LR, Edlow JA. Pitfalls in the diagnosis of cerebellar infarction. Acad Emerg Med. 2007;14(1):63-8.

12. Missed Stroke Diagnosis - John Michael Night's Story. Society to Improve Diagnosis in Medicine; 2020 [updated 2020; cited 2021 April 2]; Available from: https://www.improvediagnosis.org/stories_posts/missed-stroke-diagnosis/.

13. Kuruvilla A, Bhattacharya P, Rajamani K, Chaturvedi S. Factors associated with misdiagnosis of acute stroke in young adults. J Stroke Cerebrovasc. 2011;20(6):523-7.

14. Cano LM, Cardona P, Quesada H, Mora P, Rubio F. [Cerebellar infarction: prognosis and complications of vascular territories]. Neurologia. 2012;27(6):330-5.

15. Edlow JA, Newman-Toker DE, Savitz SI. Diagnosis and initial management of cerebellar infarction. Lancet Neurol. 2008;7(10):951-64.

16. Jauch EC, Saver JL, Adams HP, Jr., Bruno A, Connors JJ, Demaerschalk BM, Khatri P, McMullan PW, Jr.,

Qureshi AI, Rosenfield K, Scott PA, Summers DR, Wang DZ, Wintermark M, Yonas H. 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.

17. Flossmann E, Rothwell PM. Prognosis of vertebrobasilar transient ischaemic attack and minor stroke. Brain. 2003;126(Pt 9):1940-54.

18. Paul NL, Simoni M, Rothwell PM. Transient isolated brainstem symptoms preceding posterior circulation stroke: a population-based study. Lancet Neurol. 2013;12(1):65-71.

19. Hacke W, Kaste M, Bluhmki E, Brozman M, Davalos A, Guidetti D, Larrue V, Lees KR, Medeghri Z, Machnig T, Schneider D, von Kummer R, Wahlgren N, Toni D. Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. N Engl J Med. 2008;359(13):1317-29.

20. Lyden P. Thrombolytic therapy for acute stroke--not a moment to lose. N Engl J Med. 2008;359(13):1393-5.

21. Lavallee PC, Meseguer E, Abboud H, Cabrejo L, Olivot JM, Simon O, Mazighi M, Nifle C, Niclot P, Lapergue B, Klein IF, Brochet E, Steg PG, Leseche G, Labreuche J, Touboul PJ, Amarenco P. A transient ischaemic attack clinic with round-the-clock access (SOS-TIA): feasibility and effects. Lancet Neurol. 2007;6(11):953-60.

22. Rothwell PM, Giles MF, Chandratheva A, Marquardt L, Geraghty O, Redgrave JN, Lovelock CE, Binney LE, Bull LM, Cuthbertson FC, Welch SJ, Bosch S, Alexander FC, Silver LE, Gutnikov SA, Mehta Z. 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):1432-42.

23. Newman-Toker DE, Cannon LM, Stofferahn ME, Rothman RE, Hsieh YH, Zee DS. Imprecision in patient reports of dizziness symptom quality: a cross-sectional study conducted in an acute care setting. Mayo Clin Proc. 2007;82(11):1329-40.

24. Kerber KA, Brown DL, Lisabeth LD, Smith MA, Morgenstern LB. Stroke among patients with dizziness, vertigo, and imbalance in the emergency department: a population-based study. Stroke. 2006;37(10):2484-7.

25. 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.

26. Liberman AL, Newman-Toker DE. Symptom-Disease Pair Analysis of Diagnostic Error (SPADE): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data. BMJ Qual Saf. 2018;27(7):557-66.

27. Moy E, Barrett M, Coffey R, Hines AL, Newman-Toker DE. Missed diagnoses of acute myocardial infarction in the emergency department: variation by patient and facility characteristics. Diagnosis (Berl). 2015;2(1):29-40.

28. Sharp AL, Baecker A, Nassery N, Park S, Hassoon A, Lee MS, Peterson S, Pitts S, Wang Z, Zhu Y, Newman-

Toker DE. Missed acute myocardial infarction in the emergency department-standardizing measurement of misdiagnosis-related harms using the SPADE method. Diagnosis (Berl). 2020. E-pub ahead of print.

29. Nassery N, Horberg MA, Rubenstein KB, Certa JM, Watson E, Somasundaram B, Shamim E, Townsend JL, Galiatsatos P, Pitts SI, Hassoon A, Newman-Toker DE. Antecedent treat-and-release diagnoses prior to sepsis hospitalization among adult emergency department patients: a look-back analysis employing insurance claims data using Symptom-Disease Pair Analysis of Diagnostic Error (SPADE) methodology. Diagnosis (Berl). 2021. E-pub ahead of print.

30. National Quality Forum. Improving Diagnostic Quality and Safety/Reducing Diagnostic Error: Measurement Considerations [Final Report]; 2019. Contract No.: HHSM-500-2017-00060I.

31. Improving Diagnosis in Healthcare. Institute of Medicine; 2015 [updated 2015; cited 2021 April 2]; Available from: http://www.nationalacademies.org/hmd/Reports/2015/Improving-Diagnosis-in-Healthcare.aspx.

32. Naitonal Quality Forum. Advancing Chief Complaint-Based Quality Measurement [Final Report]; 2019 Contract No.: HHSM-500-2017-00060I.

33. Kerber KA, Morgenstern LB, Meurer WJ, McLaughlin T, Hall PA, Forman J, Fendrick AM, Newman-Toker DE. Nystagmus assessments documented by emergency physicians in acute dizziness presentations: a target for decision support? Acad Emerg Med. 2011;18(6):619-26.

34. 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):155-66.

35. Mane KK, Rubenstein KB, Nassery N, Sharp AL, Shamim EA, Sangha NS, Hassoon A, Fanai M, Wang Z, Newman-Toker DE. Diagnostic performance dashboards: tracking diagnostic errors using big data. BMJ Qual Saf. 2018;27(7):567-70.

36. Kerber KA, Zahuranec DB, Brown DL, Meurer WJ, Burke JF, Smith MA, Lisabeth LD, Fendrick AM, McLaughlin T, Morgenstern LB. Stroke risk after non-stroke ED dizziness presentations: A population-based cohort study. Annals of Neurology. 2014;75(6):899-907.

37. Kim AS, Fullerton HJ, Johnston SC. Risk of vascular events in emergency department patients discharged home with diagnosis of dizziness or vertigo. Ann Emerg Med. 2011;57(1):34-41.

38. 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):108-13.

39. Atzema CL, Grewal K, Lu H, Kapral MK, Kulkarni G, Austin PC. Outcomes among patients discharged from the emergency department with a diagnosis of peripheral vertigo. Ann Neurol. 2016;79(1):32-41.

40. Lee CC, Su YC, Ho HC, Hung SK, Lee MS, Chou P, Huang YS. Risk of stroke in patients hospitalized for isolated vertigo: a four-year follow-up study. Stroke. 2011;42(1):48-52.

41. Elixhauser A, Steiner C, Palmer L. Clinical Classifications Software (CCS) 2015. US Agency for Healthcare Research and Quality; 2015.

42. Kerber KA, Burke JF, Brown DL, Meurer WJ, Smith MA, Lisabeth LD, Morgenstern LB, Zahuranec DB. Does intracerebral haemorrhage mimic benign dizziness presentations? A population based study. Emergency medicine journal : EMJ. 2012;29(1):43-6.

43. Saber Tehrani AS, Kattah JC, Kerber KA, Gold DR, Zee DS, Urrutia VC, Newman-Toker DE. Diagnosing Stroke in Acute Dizziness and Vertigo: Pitfalls and Pearls. Stroke. 2018;49(3):788-95.

44. Tirschwell DL, Longstreth WT, Jr. Validating administrative data in stroke research. Stroke. 2002;33(10):2465-70.

45. Hsieh MT, Hsieh CY, Tsai TT, Wang YC, Sung SF. Performance of ICD-10-CM Diagnosis Codes for Identifying Acute Ischemic Stroke in a National Health Insurance Claims Database. Clin Epidemiol. 2020;12:1007-13.

46. 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.

47. Zhou Y, Lee SH, Saber Tehrani AS, Robinson KA, Newman-Toker DE. Anterior Circulation Stroke Causing Dizziness or Vertigo: A Systematic Review [abstract]. 136th Annual Meeting of the American Neurological Association; September 25–27, 2011; San Diego, CA.

48. 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; November 10-13, 2019; Washington, DC.

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

***If a COMPOSITE** (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.*

Diagnostic error is a major public health problem.[1] The lack of operational measures is a critical barrier to improving diagnosis.[2,3] 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.[4] Among vascular events, missed stroke is the leading cause of serious harm to patients. Misdiagnosis of stroke disproportionately occurs when symptoms and signs are not typical or obvious.[5,6] Among strokes, the most

commonly missed clinical presentation is patients presenting dizziness or vertigo, easily mistaken for inner ear disease.[5] In US emergency departments (ED) each year, an estimated 45,000-75,000 patients with strokes presenting dizziness or vertigo are missed and erroneously discharged.[6]

ED patients with acute dizziness and vertigo could be diagnosed correctly using evidence-based bedside examinations,[7,8] but there is currently a large evidence-practice gap[9] in ED diagnosis, resulting in substantial harms to patients.[6] Without timely, accurate diagnosis, these patients suffer misdiagnosis-related harms[10] from lack of prompt treatment.[5] The most common harm is preventable major stroke after minor stroke or transient ischemic attack (TIA),[11,12] with major stroke leading to subsequent hospitalization. Crude short-term stroke hospitalization rates per 10,000 dizziness discharges from the ED vary at least from 20-80.[6] Adjusting for baseline stroke risk across groups does not eliminate practice variation.[13]

This outcome measure tracks the rate of patients admitted to the hospital for a stroke within 30 days of being treated and released from the 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”). This measure is the first operationally-viable performance measure of stroke misdiagnosis for the hospital setting. Hospital EDs will be able to use the measure internally to track their performance over time, as they work to implement interventions to reduce misdiagnosis of strokes. The measure can also be used by external entities for public reporting and pay-for-performance, as external pressures to encourage hospital improvements.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis*. (**This is required for maintenance of endorsement**. Include mean, std dev, min, max, interquartile range, 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 (4b1) 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.3. If no or limited performance data on the measure as specified is reported in 1b2, 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.

1b.4. 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. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) 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 (4b1) under Usability and Use.*

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, 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 1b.4

Prior research has identified that women and minorities are at ~20-30% increased odds of stroke misdiagnosis and patients 18-44 years old[5] are at roughly 7-fold increased odds.[14,15]

Differences by Gender

• Newman-Toker et al., Diagnosis, 2014: Found the odds of a probable misdiagnosis were lower among men (OR 0.75) than women.

• Von Kleist et al., Neurology, 2019: 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: 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: Found patients age <=35 years (P=.05) were more likely to be misdiagnosed.

• Newman-Toker et al., Diagnosis, 2014: 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).

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*

De.6. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

https://dxcenter.org/AvoidHARM-Dx-Measures/Dizzy-Stroke-ED

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)

This is not an eMeasure **Attachment:**

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

Attachment**:** ICD_codes.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

S.4. Numerator Statement *(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.*

***IF an OUTCOME MEASURE**, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).*

The number of ED index visits during the performance period that are followed within 30 days by an inpatient hospital admission to any hospital that ends in a primary diagnosis of stroke.

S.5. Numerator Details *(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 S.2b)*

***IF an OUTCOME MEASURE,** describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).*

For each patient's ED index 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-9 and ICD-10 codes to be used to identify patients with a primary diagnosis of stroke can be found in the submitted Excel file.

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

Patients discharged from the ED with "benign dizziness" as the primary diagnosis code, counting a patient's first such discharge during the performance period (an "index visit) and all subsequent such discharges that fall outside a 360-day follow-up window from the previous qualifying "index visit".

S.7. Denominator Details *(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 S.2b.)*

***IF an OUTCOME MEASURE**, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).*

Using a 36- month performance period, identify those ED patients who were discharged from the ED with a primary 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-9 and 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 discharge during the performance period meeting the above criteria should be included in the denominator. This patient discharge is considered the patient's first "index visit". A patient's second "index visit" is the first subsequent ED 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 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 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.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

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

S.9. Denominator Exclusion Details *(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 S.2b.)*

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including 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 with at S.2b.)*

Not applicable.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Lower score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

S.12. Type of score:

Rate/proportion

If other:

S.13. Interpretation of Score (Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)

Better quality = Lower score

S.14. Calculation Algorithm/Measure Logic (Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)

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

a. Step 1 – Identify all patients discharged from the ED with a primary 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 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 discharge with a 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 discharge with a 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 discharge with a 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 were excluded.

c. Step 3 – Count the number of ED "index visits". This is the denominator value.

"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 discharge that resulted in a primary diagnosis of stroke. Count the number of "index visits" that meet this criterion. This is the immediate 30-day numerator value.

e. Step 5 – Measure the observed rate. Crude immediate 30-day rate per 10,000 visits = (Step 4: [number of immediate 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 in the time window 91 days through 360 days following their ED index visit discharge that resulted in a primary diagnosis of stroke. Count the number of "index visits" that meet this criterion.

g. Step 7 – Divide the number of 91 days through 360 days strokes identified in Step 6 by 9 to calculate the average delayed rate per 30 days. This is the delayed 30-day numerator value.

h. Step 8 – Measure the expected rate. Crude delayed 30-day rate per 10,000 visits = (Step 7: [average number of delayed stroke hospitalizations per 30d + alpha] / Step 4: [number of eligible ED benign dizziness 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 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 immediate 30d rate per 10,000 visits = Step 5 (crude immediate 30d rate) – Step 8 (crude delayed 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 discharge 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.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

Not applicable.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, 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.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Claims

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

The measure is calculated from claims data. If the ED discharge status is known to be coded poorly in available claims data, additional data cleaning using electronic health records may be required.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Emergency Department and Services

If other:

S.22. COMPOSITE Performance Measure - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2. Validity – See attached Measure Testing Submission Form

NQF_Avoid_Harm_Dizzy-Stroke_Measure_Testing_210409.docx

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include

information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:**
Measure Title: **Hospitalization After ED Release with Missed Dizzy-Stroke (Avoid H.A.R.M. Dizzy-Stroke)**
Date of Submission: **1/5/2021**

Type of Measure:

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

*cell intentionally left blank

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

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

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

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

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g.,*

Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).

Calculating and testing the performance measure: 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.

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).

1.3. What are the dates of the data used in testing? 07/01/2014 -12/31/2017

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

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

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

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

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

ED Visits per year	Muelleman et al. (2007 data) N=4,874 Non-Federal EDs	Our Data Set (2017 data) N=5,503 Medicare EDs
<10,000	31%	32%
10,000-19,999	21%	16%
20,000-29,000	15%	12%
30,000-39,999	13%	10%
40,000-49,000	8%	8%
>50,000	12%	23%

For the actual measure analysis, we used 967 of the 5,503 facilities. These 967 facilities had at least 250 "benign dizziness" discharges from the ED during the 3-year performance period and therefore would have a large enough sample size to produce a reliable measure of performance. Hospitals with 250 "dizzy out" discharges in Medicare data over 3 years typically reflect a hospital ED that sees 40,000 to 50,000 ED visits per year (depending on the proportion of patients at that ED using Medicare insurance).

Due to data privacy constraints, we were not able access descriptive statistics of the 967 facilities used in the actual measure analysis. But these 967 facilities would likely include those U.S. hospital EDs that are in the top 20-25% of annual ED visits. Besides the logical characteristics of larger EDs (i.e., located in larger population centers), we would not anticipate any additional systematic biases in the facilities that were included in the analysis.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

For testing the measure, a total of 1,232,389 patients with a "dizzy" discharge diagnosis were included in the testing and analysis. These reflect patient discharges from the 967 hospital EDs during the 3 year performance period.

Data from 2015 to 2017: **Patient Demographics of ED Discharges with a "Dizzy" Diagnosis**	Data from 2015 to 2017: **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%

Data from 2015 to 2017: **Patient Demographics of ED Discharges with a "Dizzy" Diagnosis**	Data from 2015 to 2017: **Percentage of Patients (%)**
• 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%

*cell intentionally left blank

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

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

Data-element validity testing (Johns Hopkins Health System): Data from July 1, 2014 – June 30, 2015 (used to test ICD-9) and July1, 2016 – June 30, 2017 (used to test ICD-10) were used for data-element validity testing.

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

No social risk factors were analyzed. However, our "observed minus expected" approach that accounts for baseline stroke risk does account for social determinants of long-term stroke risk in the cohort of patients who are at risk and being measured.

2a2. RELIABILITY TESTING

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

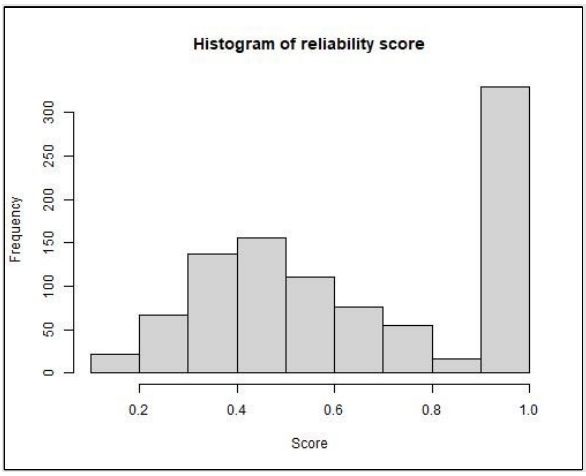
2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

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

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

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



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 "dizzy out" discharges in the three year performance window to look at the median reliability score for each stratum.

Number of "Dizzy Out" Discharges in the 3 Year Performance Window	Median Reliability Score
250-499	0.582
500-749	0.710
750+	0.807

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

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 providers) whereas a reliability of 1.0 implies that all variation is caused by real difference in performance across accountable entities. The reliability score is

dependent upon the pool of facilities that are included in the sample and the reliability score is unique to each facility in that pool.

And while there is not a clear cut-off for a minimum reliability level, a median value very close to 0.60 (0.590) would be considered by many to be sufficient for seeing differences between some facilities. This finding holds true even for the smallest facilities included in the analysis (those with 250-499 "dizzy out" discharges over 3 years), where we saw a median reliability score value of 0.582.

Due to the low H.A.R.M. event rates being measured (which reflect clinical adverse events after a missed diagnosis of stroke), it is anticipated that high reliability would require a higher number of visits (e.g., >500 dizziness discharges over 3 years). This value corresponds to a hospital ED with ~15,000 visits per year (any complaint). The majority of EDs in the US have visit volumes over this threshold. It is only because Medicare data reflect a ~20% sample of visits that the 3-year measure provides high reliability only for larger institutions. In other words, other data sources such as EHR, regional health information exchange, or state-level (e.g., HCUP) data that have complete case capture could be used with this measure to assess diagnostic performance at very high reliability, assuming similar results hold.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☐ **Empirical validity testing**

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

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

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

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

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

2. McCormick et al. (*PLoS One*, 2015; PMID: 26292280) conducted a systematic review of studies reporting on the validity of International Classification of Diseases (ICD) codes for identifying stroke in administrative data.

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

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

Measure denominator (patients discharged from the ED with a diagnosis of benign dizziness)

PART A

For dizziness (denominator = ED dizziness 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 code, how often do charts suggest the ED provider INTENDED to code a "benign dizziness" discharge?

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:

- ICD-9-CM: Jul 2014 – Jun 2015
- ICD-10-CM: Jul 2016 – Jun 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 match 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-9/10-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 64 charts for each non-dizziness sub-group and for each performance period, to estimate the positive predictive value (PPV) of the benign dizziness discharge codes. For performance period Jul 2014 – Jun 2015 (ICD-9 period), 3 hospitals were included (Hospitals A, C, D; Hospital B had not yet transitioned to the Epic electronic health record) in the stratified sampling. For performance period Jul 2016 – Jun 2017 (ICD-10 period), all 4 hospitals were included in the stratified sampling.

PART B

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

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:

- ICD-9-CM: Jul 2014 – Jun 2015
- ICD-10-CM: Jul 2016 – Jun 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 criteria --- "a OR b OR c")
  - a) ED (Epic) chief complaint of dizziness/vertigo at ED Index Visit
  - b) Benign dizziness diagnosis (HCUP CCS 6.8.2) in a non-primary position at ED Index Visit
  - c) Middle 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, as their presenting symptoms match their discharge diagnosis. We did not review these charts manually. For the high-risk for misclassification group, we manually reviewed charted records 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-9/10-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 high-risk sub-group, to estimate the negative predictive value (NPV) of the "not benign dizziness" discharge codes. For performance period Jul 2014 – Jun 2015 (ICD9 period), 3 hospitals were included (Hospitals A, C, D; Hospital B had not yet transitioned to the Epic electronic health record) in the stratified sampling. For performance period Jul 2016 – Jun 2017 (ICD10 period), all 4 hospitals were included in the stratified sampling.

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 June 2013 and June 2018. 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, given that 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.

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

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

Key results from the articles mentioned above are as follows:

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

2. The McCormick systematic review (PMID: 26292280) included 77 published 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 stroke 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. In the Kokotailo and Hill study (PMID: 16020772) they found that stroke coding was equally good with ICD-9 (90% correct [95% CI 86-93]) and ICD-10 [92% correct (95% CI 88-95). There were some differences in coding by stroke type, notably with transient ischemic attack, but these differences were not statistically significant.

Measure denominator (patients discharged from the ED with a diagnosis of benign dizziness)

PART A

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

Performance Period and CC categories	Number of ED Index Visits	Number of matched records	Proportion estimates of matched records	95% confidence intervals
JHHS – Jul 2014 – Jun 2015 (testing ICD-9)	1483	*	*	*
CC dizziness	1052	1052/1052*	100%*	99.65-100%*
CC oto-vestibular	105	32/32	100%	89.11-100%
CC other	326	32/32	100%	89.11-100%
JHHS – Jul 2016 – Jun 2017 (testing ICD-10)	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%
GRAND TOTAL	3309	2488/2488	100%	99.89-100%

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

*cell intentionally left blank

PART B

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

Performance Period and risk categories	Number of ED Index Visits	Number of matched records	Proportion estimates of matched records	95% confidence intervals
JHHS – Jul 2014 – Jun 2015 (testing ICD-9)	76007	*	*	*
High risk group	1761	62/67	92.54%	83.44-97.53%
Low risk group	74246	74246/74246*	100%*	99.995-100%*

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 (testing ICD-10)	99464	*	*	*
High risk group	12744	66/67	98.51%	91.96-99.96%
Low risk group	86720	86720/86720*	100%*	99.996-100%*
GRAND TOTAL	175471	161094/161088	99.997%	99.993-99.999%

** These charts were not manually reviewed, but 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.*

*cell intentionally left blank

DISCHARGE STATUS

100% of the 25 ED patient 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 of the measure were found to have accurate status (3 were discharged from the ED to observation/clinical decision units and returning with stroke hospitalizations Days #1-30; 3 were day #1 return hospitalizations).

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

Measure Numerator

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

Measure Denominator

We found a positive predictive value (PPV) of 100% [CI: 99.89%-100.00%] for coding "benign dizziness." Of the 128 charts reviewed (and 3,181 electronically confirmed), all of the patients coded with a "benign dizziness" diagnosis at discharge 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 great confidence that the codes we have outlined for identifying "benign dizziness" patients are indeed capturing patients where the provider intended that diagnosis.

We found a negative predictive value (NPV) of 99.997% [CI: 99.993-99.999%] for coding "not benign dizziness." Of the 134 charts reviewed (and 175,331 electronically confirmed), all but 6 of the charts that were coded as "not benign dizziness" diagnosis at discharge 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. There were 6 charts among the oversampled high-risk group that did not have a primary diagnosis of "benign dizziness" at discharge, but the charted record suggested the ED provider intended to assign a "benign dizziness" code. Even among the high-risk subset, the vast majority of cases were

confirmed 95.52% [CI: 90.51-98.34]. Given the high NPV (99.9%+) that was found, we feel confident that the coding is clearly 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 when assessing all the highest-risk cases.

2b2. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

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

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

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

_____

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

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

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

N/A

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

Our measure uses a statistical risk difference approach (observed [short-term stroke risk] minus expected

[long-term/baseline stroke risk]). As a result, controlling for differences in patients characteristics (case mix) is not needed to achieve fair comparisons across entities. .

Risk Difference Approach: The risk-difference measure is a difference between two rates (observed minus expected), reflecting the observed stroke events in the first 30 days that have occurred above the expected epidemiologic base rate (i.e., are likely to represent more than a chance association between the ED discharge and inpatient admission). This approach accounts for inter-institutional differences in the underlying stroke risk of their specific patient populations, including any social determinants of 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., >90 days post discharge) are not potentially 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 ***immediate 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 delayed outcome assessment window. The delayed window (91 days to 360 days post discharge) is chosen to reflect the epidemiologic base rate of stroke (i.e., after the immediate, short-term risk of a misdiagnosis leading to preventable major stroke following minor stroke has definitively passed). The stroke rate per 30-day period during this delayed 270-day window is obtained by dividing the numerator by nine and is called the ***delayed 30-day rate***.

Risk Difference Rationale: Patients that have stroke hospitalizations within 30 days of an ED "benign dizziness" discharge consist mostly of 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. 1). This curve matches the natural history biological profile of major stroke following minor stroke and TIA (Fig. 2A/B).

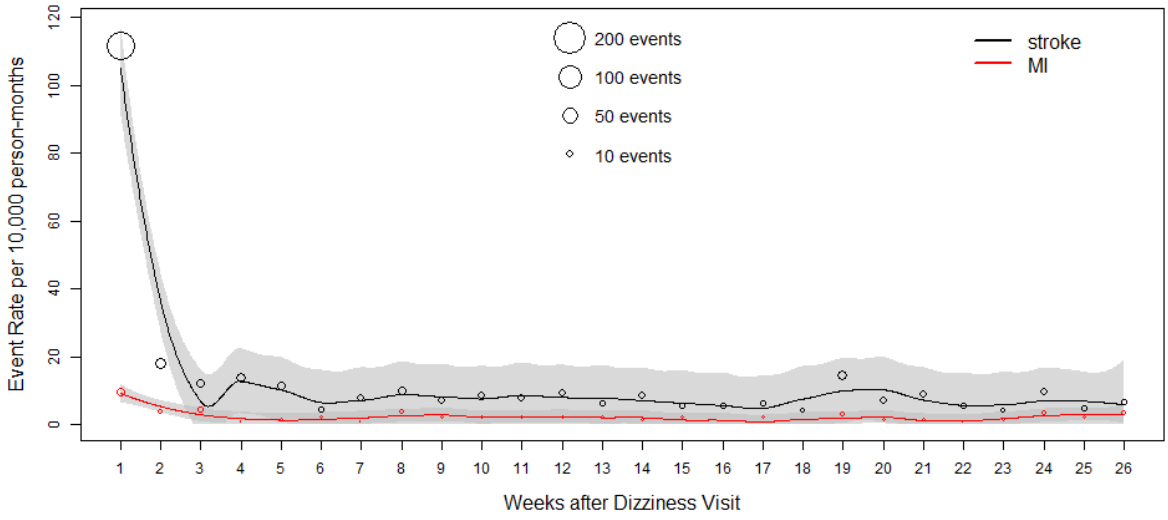
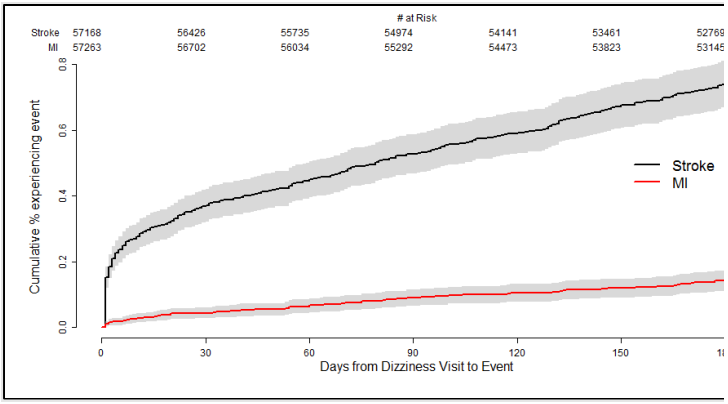


Figure 1. Weekly incidence rate curve of stroke hospitalization post ED discharge as "benign dizziness." Kaiser Permanente Mid-Atlantic data from the performance period from 2010-2014 at all outpatient sites (ED, ambulatory care). Data reflect 56,746 treat-and-release visits for "benign dizziness." Shown in black are stroke hospitalizations, and shown in red are heart attack hospitalizations. 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).



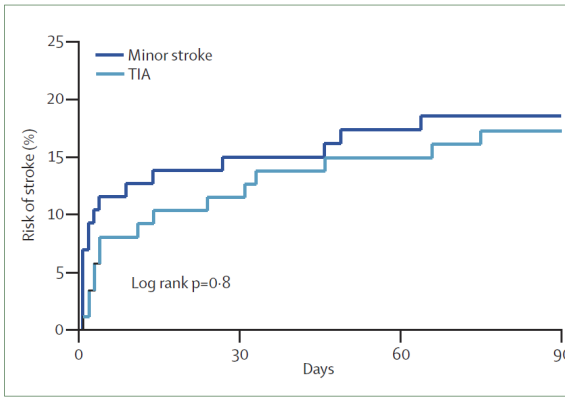


Figure 2A. Cumulative incidence curve of stroke hospitalizations post ED discharge as "benign dizziness." Represented here are the same data as shown in Figure 1. These data are presented here as a cumulative incidence curve for comparison to Figure 2B at right.

Figure 2B. Cumulative incidence curve for major stroke following TIA or minor stroke. Data are from the Oxford Vascular Study as represented in Rothwell, Buchan, & Johnston (*Lancet Neurol*, 2006; PMID: 16545749).

This risk difference approach uses an institution-specific delayed (91d–360d) stroke hospitalization rate to approximate the baseline short-term stroke risk for the population in question. This delayed 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 2B). 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 incorrect diagnosis that are linked to the institution-specific patient population should be measured appropriately, rather than "adjusted" away (e.g., racial bias, which leads racial minorities to be at higher risk of being misdiagnosed [Newman-Toker et al., *Diagnosis*, 2014; PMID: 28344918]).

Risk Difference Calculation: The risk difference calculation requires an observed and expected rate calculation. For each patient discharged from the ED with a "benign dizziness" diagnosis during the performance period, data on stroke hospitalizations must be available for a floating outcome assessment window of roughly 12 months (360 days). If stroke hospitalizations occur between post-ED day #1 and day #30 (i.e., mostly linked to misdiagnosis-related harms), they are counted in the numerator of the "immediate 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 "delayed 30-day rate." The delayed 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** delayed 30-d rate calculation because of very low stroke base rates in this population (<0.1% [Newman-Toker, *Ann Neurol*, 2015; PMID: 26418192]); this increases the precision of the "expected" value.

- **Crude immediate 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 "immediate" 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 delayed 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 who did not experience a stroke in the prior 90 days + 1 - (3 x *alpha*)]} x 10,000. This "delayed" rate approximates the epidemiologic "base" rate of stroke in the specific population in whom the immediate 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, as 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 immediate 30d rate** = (crude immediate 30d rate) – (crude delayed 30d rate); the attributable immediate 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.

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

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 incorrect diagnosis that are linked to the institution-specific patient population should be measured appropriately, rather than "adjusted" away (e.g., racial bias, which may place minorities at higher risk of being misdiagnosed [PMID: 28344918]).

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

N/A

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

N/A

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

N/A

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

used)
N/A

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

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

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

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
N/A

2b3.9. Results of Risk Stratification Analysis:

N/A

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

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

N/A

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

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

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

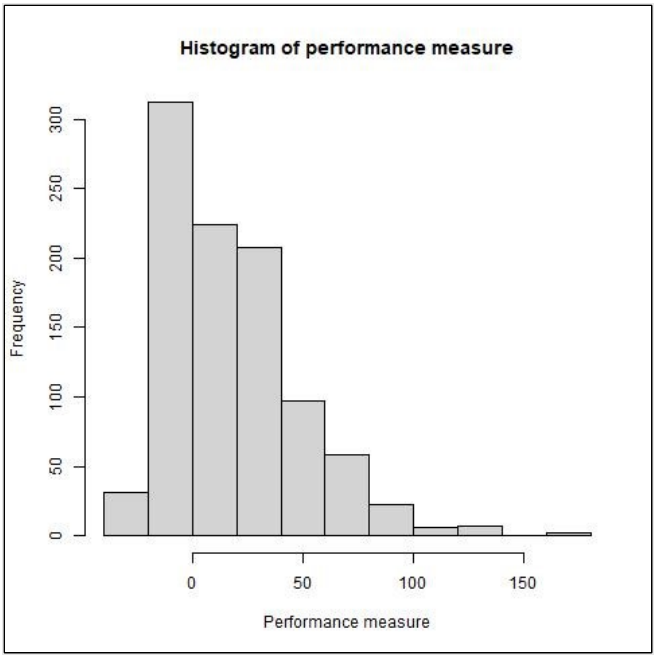
Our first strategy was to calculate common descriptive statistics that would help summarize the distribution of performance scores to see if there is meaningful variation across facilities. This included calculating the mean, median, standard deviation, and interquartile range of all the of the facility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

Our second strategy was to calculate a 95% confidence interval around each facility's score and whether the confidence interval included the national average. If the confidence interval did not include the national average, the facility is then identified as being better or worse than average. We also looked to see 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).

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

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



Attributable 30-day Stroke Misdiagnosis 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

- 627 of 967 hospitals were identified as being "Better" than the national average (upper bound of 95% CI was less than national average)
- 0 of 967 hospitals were identified as being "Worse" than the national average (lower bound of 95% CI was greater than national average)
- 8 of 967 hospitals were identified as having statistically significant "Harm" (lower bound of 95% CI was greater than zero)

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

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 25th percentile is close to the difference between the median and the 75th percentile).

With the measure we were also 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".

As described above in **2a2.4**, we expect that this same measure, used in clinical practice with a more complete data set reflecting all ED dizziness discharges (rather than only the 20% Medicare fraction), would demonstrate even greater precision to identify differences among facilities.

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

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

N/A

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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.

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

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

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

3. Feasibility

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

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in electronic claims

3b.2. 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 other than electronic sources. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. Required for maintenance of endorsement. 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.

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

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-9-CM or 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 is calculated using Medicare claims 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 using regional health information exchanges or at the state level using curated data sets us as AHRQ's HCUP SID and SEDD data, for states where linkable data sets are available (at least 14 states currently have such capabilities https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6929527/).

The main tradeoff when using Medicare data for national benchmarking is that Medicare data represent only ~20% of the sample of patients with dizziness 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 meaningful way for hospitals to track their own internal performance, while their official performance is calculated from stroke returns to all hospitals. 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 NQF 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.

3c.2. Describe 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)*.

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.

4. Usability and Use

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

4a. 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 not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Public Health/Disease Surveillance Payment Program Quality Improvement (external benchmarking to organizations) Quality Improvement (Internal to the specific organization)	*

*cell intentionally left blank

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

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

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This is a newly developed measure, so it is currently not being publicly reported or being used in an existing accountability program.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*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 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 smaller 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 NQF, 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 an NQF 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 an NQF 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 an NQF 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 an NQF measure endorsement.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

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 (PMID: 29550767).

4a2.1.2. Describe the process(es) involved, 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 NQF. 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).

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Based on these measures and the success of the Tele-Dizzy program (see 1a.2 TELE-DIZZY CONSULTATION SERVICE), Johns Hopkins has agreed to extend its stroke reduction initiatives to other hospitals within the Johns Hopkins Health System. 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 planning to submit a grant proposal 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).

4a2.2.2. 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) 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." 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 now 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]). Additional stakeholder feedback will be forthcoming as part of R01 HS 27614 over the years 2021-2024.

4a2.2.3. 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 (Improving Diagnostic Quality and Safety, 2017; Advancing Chief Complaint-Based Quality Measurement, 2019; Improving Diagnostic Quality and Safety/Reducing Diagnostic Error: Measurement Considerations, 2020).

4a2.3. Describe how the feedback described in 4a2.2.1 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 analysis presented here.

Furthermore, 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.

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. Refer to data provided in 1b 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, what are the reasons? 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 (PMID: 29482196). 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 at least partially supported by our analysis showing that larger hospitals are outperforming smaller hospitals (see 1a.2, large vs. small hospitals) combined with imaging use trends. 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% (PMID: 21570240) and some large academic centers having current MRI rates of up to 20% (Gold et al., Diagnostic Error in Medicine, 2019). This again reinforces the need for balance measures, as noted in 4a2.3.

4b2. Unintended Consequences

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).

4b2.1. Please 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.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

N/A

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

No

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** NQF_Appendix_3614.pdf

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Johns Hopkins Armstrong Institute for Patient Safety and Quality

Co.2 Point of Contact: Matt, Austin, jausti17@jhu.edu, 832-816-5618-

Co.3 Measure Developer if different from Measure Steward: Johns Hopkins Armstrong Institute for Patient Safety and Quality

Co.4 Point of Contact: Matt, Austin, jausti17@jhu.edu, 832-816-5618-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2021

Ad.3 Month and Year of most recent revision: 04, 2021

Ad.4 What is your frequency for review/update of this measure?

Ad.5 When is the next scheduled review/update for this measure?

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: