Partnership for Quality Measurement
Powered by Battelle

# Fall 2023 Endorsement and Maintenance (E&M) Committee Independent Review Summary

## INITIAL RECOGNITION AND MANAGEMENT COMMITTEE

**Prepared by:**

**Battelle**
**505 King Avenue**
**Columbus, Ohio 43201**
**January 2024**

The analyses upon which this publication is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services.

*image by rawpixel.com*

# Table of Contents

# Summary of Committee Independent Reviews

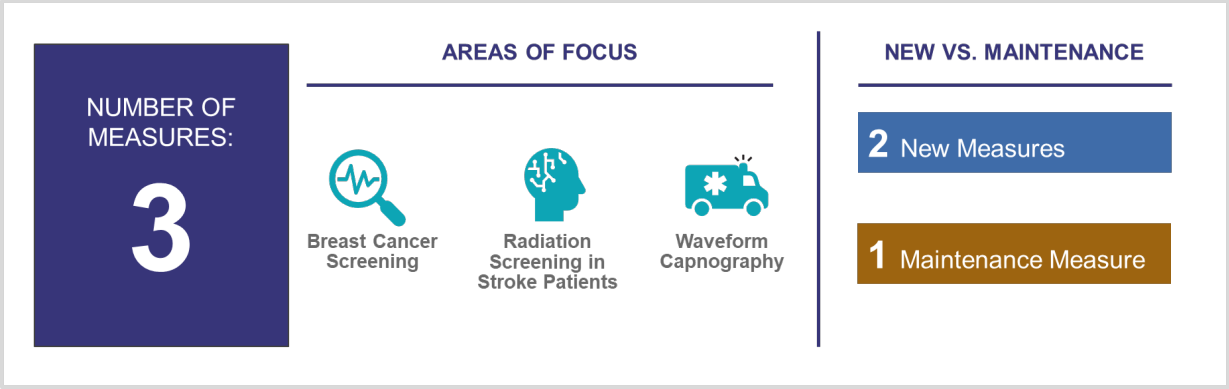
## Independent E&M Committee Member Reviews Overview

At least three (3) weeks prior to an E&M committee endorsement meeting, the Recommendations Group and the Advisory Group of each E&M committee receive the full measure submission details for each measure up for review, including all attachments, the Partnership for Quality Measurement (PQM) Measure Evaluation Rubric, the public comments received for the measure(s) under review, and the E&M team preliminary assessments.

Members of both groups were asked to review each measure, independently, against the PQM Measure Evaluation Rubric. Committee members assigned a rating of "Met," "Not Met but Addressable," or "Not Met" for each domain of the PQM Measure Evaluation Rubric. In addition, committee members provided associated rationale for each domain rating, which is based on the rating criteria listed for each domain. Battelle staff aggregated and summarized the results and distributed them back to the committee, and to the respective measure developers and/or stewards, for review within one (1) week of the endorsement meeting.

These independent committee member ratings are compiled and used by Battelle facilitators and committee co-chairs to guide committee discussions.

*Figure 1. Fall 2023 Measures for Committee Review*




For the Fall 2023 cycle, the Initial Recognition and Management committee received three (3) measures, two (2) new measures, and one (1) measure undergoing maintenance endorsement review (Figure 1). The measures focused on breast cancer screening, radiation screening in stroke patients and waveform capnography.

# Measure-Specific Summaries

The following brief summaries include themes and considerations gathered from the committee's independent reviews for each of the five domains of the PQM Measure Evaluation Rubric. Themes were assessed and categorized with respect to the strengths and limitations of the measure(s) under endorsement review. Corresponding to the themes are the number of committee reviews received and stratified by the ratings of "Met," "Not Met," and "Not Met, but Addressable."

## CBE #4220 - Breast Cancer Screening Recall Rates

**Number of Committee Reviews:** 24

| Importance (n=24) | Strengths | Limitations |
|---|---|---|
| **Consensus**<br><br>17% Met<br><br>83% Not Met, but Addressable<br><br>0% Not Met | • **Importance of the Measure:** The measure is deemed important for preventative and diagnostic mammograms. It's seen as a balancing measure to tracking rates of screening alone.<br><br>• The information provided supports the importance of the measure. It's particularly important as it focuses on monitoring "rates of recall following screening imaging instead of rates of breast cancer imaging."<br><br>• This is a new measure under PQM, but it is a re-specification measure for the "Mammography Follow Up Rates" under OQR.<br><br>• **Benefits of Screening and Follow-up:** There is strong evidence presented about the benefits of screening for breast cancer and follow-up evaluations when needed. It is further presented that there are negative consequences when mammography and digital breast tomosynthesis (DBT) recall rate is either too high or too low. | • **Recall Rate Range:** Questions are raised about the rationale for the 5%-12% range. There's a suggestion to consider if current conditions warrant a shift in the range or having different ranges for special circumstances.<br><br>• **Patient Engagement:** There's a concern about limited patient engagement on the importance of the measure.<br><br>• **Improved Outcomes:** Unclear how this measure will support improved outcomes. Connection between this measure and reduction of radiation is unclear.<br><br>• The current measure may not adequately address the need for appropriate clinical follow-up without overuse. The impact of adding DBT on the ACR screening target recall rate wasn't addressed.<br><br>• **Quality Improvement Strategies:** The measure does not address how the screening recall rate can be used to improve performance.<br><br>• **Performance Gaps:** While gaps are reported in deciles with 4.9% and 13.0% as the closest deciles, this does not |

| Importance (n=24) | Strengths | Limitations |
|---|---|---|
| | • The developer cites evidence and guidelines from the American College of Radiology. | allow precise assessment of the gap from the 5-12% suggested. In addition, the high recall rate category (over 13%) has a performance level mean of 13.7% which demonstrates a measurable but not clearly meaningful performance gap in over utilization based on consensus guidelines. |

| Feasibility (n=24) | Strengths | Limitations |
|---|---|---|
| **Consensus**<br>92% Met<br>8% Not Met, but Addressable<br>0% Not Met | • **Feasibility and Data Collection:** The measure is feasible and does not present an undue burden on hospitals for data collection.<br>• The data required is routinely generated from patient encounters, claims, and the EHR utilizing value sets. The measure is already being used in the CMS Outpatient Imaging Efficiency program.<br>• The burden of reporting this measure was directly addressed by the developer, and no proprietary data is needed. | • **Feasibility and Data Collection:** There were some concerns about the representativeness of the sample used to evaluate feasibility. The geographic characteristics of the individuals were not reported, which could be important as professionals in rural areas may report an undue burden with collecting this data.<br>• **Patient Representation:** Developer evaluated feasibility amongst 32 individuals. However, the patient representation relative to health care staff/professionals could undermine the feasibility assessment result. It would be appropriate for the developers to consider a patient-only group to ensure accurate representation. |

| Scientific Acceptability (n=24) | Strengths | Limitations |
|---|---|---|
| **No Consensus**<br><br>71% Met<br><br>29% Not Met, but Addressable<br><br>0% Not Met | • **Reliability:** Most facilities exceeded the accepted threshold of 0.6. The median reliability score was 0.95.<br><br>• **Data Adequacy:** The data used for testing were adequate, with good representation between facilities/patients. The measure uses the same data process as Outpatient Imaging Efficiency measures using claims data.<br><br>• **Measure Design:** The measure specifications are well-defined and precise. The measure as outlined was well designed.<br><br>• **Validity:** The developer facilitated qualitative assessments of the measure's validity. Face validity results are acceptable. Validity testing was completed and consensus was reached. | • **Outliers:** The developers did not describe the reliability of 0.41, which is a steep difference compared to the mean of 0.9 and the next lowest value of 0.81.<br>• **Risk Adjustment:** The measure is not risk-adjusted as it is a process measure. This could impact recall rates, potentially leading to unnecessary exposure to radiation or follow-up testing in low-risk populations.<br><br>• If there is a high risk of cancer in the population then the appropriate recall rate may be higher and even outside the suggested 5-12% range. Similarly, in a low-risk population, not adjusting the pass rate for the measure would increase the exposure to radiation or follow up testing (biopsy) unnecessarily.<br><br>• **Validity Methods:** The validity methods rely on a consensus of 32 individuals, but it's unclear if they represent broad stakeholders in breast cancer care.<br><br>• **Inclusions/Exclusions:** There are concerns about including men in the denominator, since they are typically screened only if they have significant risk factors or symptoms. There are concerns about the rationale for removing MRI as a follow-up imaging modality. |

| Equity (n=24) | Strengths | Limitations |
|---|---|---|
| **Consensus**<br>92% Met<br>4% Not Met, but Addressable<br>4% Not Met | • **Equity Assessment:** The developers conducted a thorough assessment of equity for sex, race/ethnicity, age, and dual eligibility status. They used performance data to calculate the rate of recall by these factors and found overall significance. | • **Limited Scope:** There are concerns that the Medicare FFS measure will not capture many of the patients in populations at risk for receiving inequitable recall rates. It would have been beneficial to see the ethnic/race data based on rural and urban settings and size of facilities.<br>• **Clinical Significance:** While statistical testing indicates a statistically significant difference in recall rates, there is no indication if this is clinically significant, appropriate based on risk, or equitable/inequitable.<br>• **Addressing Differences:** The measure developer can establish a difference in the measure across different patient groups. However, it does not clearly establish how the measure supports addressing these differences. |

| Use and Usability (n=24) | Strengths | Limitations |
|---|---|---|
| **No Consensus**<br>21% Met<br>67% Not Met, but Addressable<br>13% Not Met | • **Measure Performance and Improvement:** The measure is seen as a useful tool for internal quality measurement and can help identify opportunities for improvement.<br>• Most of the multi-stakeholder group agreed the measure could be used by entities for QI and decision-making (77.4%) and that it would provide consumers and providers with actionable information (80.6%). | • **Measure Performance and Improvement**: The developer did not address how an organization could take actions to improve performance if rates fell outside of the 5% and 12% parameters.<br>• There's uncertainty about what follow-up action medical facilities should take when their scores are higher than expected. |

| Use and Usability (n=24) | Strengths | Limitations |
|---|---|---|
| | • **Patient Care and Outcomes**: The measure could potentially increase use and usability, which is important for patients who have abnormal mammograms and often wait long for further testing.<br>• **Pay for Performance Programs:** Developer plans for the measure to be used in CMS's Hospital Outpatient Quality Reporting (HOQR) Program, a pay-for-quality program. | • **Patient Care and Outcomes**: There's concern about the interpretation of results for facilities with too low or too high rates. It's unclear whether patients will use this data.<br>• **Pay for Performance Programs**: The measure is seen as inappropriate for pay for performance programs due to the potential for unintended consequences. The use of a "range" in pay for performance could incentivize inappropriate actions.<br>• **Variability and Differences**: There's concern that detection of variability or clusters due to geographic, cultural, or socioeconomic factors could be discouraged or misinterpreted. Concerned about facilities in underserved areas. |

## CBE #0661 - Head CT or MRI Scan Results for Acute Ischemic Stroke or Hemorrhagic Stroke Patients who Received Head CT or MRI Scan Interpretation within 45 minutes of ED Arrival

**Number of Committee Reviews:** 21

| Importance (n=21) | Strengths | Limitations |
|---|---|---|
| **Consensus**<br>86% Met<br>14% Not Met, but Addressable<br>0% Not Met | • Strong evidence for this measure for the care of suspected stroke patients in the ED with multiple studies and clinical guidelines showing the importance of timely imaging and intervention in acute ischemic, supporting the measure.<br>• Comments emphasize the importance of early stroke identification and treatment, with questions about potential barriers to rapid access to scanning.<br>• Patients surveyed find the measure valuable. | • Measure performance has stagnated, with no improvement seen in the last five years.<br>• The developer presents the existing quality measure including measure characteristics and specifications. While this has a high level of detail, it does not outline the importance of this measure.<br>• The description of patient input does not support the conclusion that time-to-interpretation is meaningful for patients.<br>• What is the rationale for the 45-minute time limit for the CT scan or MRI, as it is unable to find it in the literature cited |

| Feasibility (n=21) | Strengths | Limitations |
|---|---|---|
| **Consensus**<br>86% Met<br>10% Not Met, but Addressable | • Data for the measure are generated during care and uses data from EHRs or other electronic sources.<br>• The measure is currently being implemented and has been in use for a long time, demonstrating feasibility. | • The measure requires chart abstraction to report as specified, which may be a significant fee or cost.<br>• If a provider/facility does not have software to do the abstraction (which could be expensive), they will need |

| Feasibility (n=21) | Strengths | Limitations |
|---|---|---|
| 5% Not Met | • The measure appears to be feasible as data elements are collected in the normal course of care. | manual extraction, which can also be expensive and time consuming. |

| Scientific Acceptability (n=21) | Strengths | Limitations |
|---|---|---|
| **No Consensus**<br>43% Met<br>57% Not Met, but Addressable<br>0% Not Met | • The measure is well-defined, and the data are stable without improvement.<br>• The specifications are clear, and the reliability results are fair overall.<br>• The measure is suitable and accurately specified.<br>• Data element validity is clearly established. The validity testing results were reassuring that validity is adequate and specific threats to validity weren't identified. | • There are concerns about the reliability testing, with approximately 30-35% of entities having reliability less than 0.6.<br>• Several suggestions for improvement are made, including increasing the minimum case volume, extending the timeframe for the measure, and considering a mitigation strategy for facilities with a low denominator.<br>• There are concerns about the hypothesis testing confirming a difference between before performance for male vs female patients.<br>• Some would have preferred a different approach to empirical validity of the measure score.<br>• There's a suggestion to know more about any issues with data element "Head CT/MRI Scan Interpretation Time". |

| Equity (n=21) | Strengths | Limitations |
|---|---|---|
| **Consensus**<br>10% Met<br>10% Not Met, but Addressable<br>81% Not Met | • None | • Developer did not address this optional criterion.<br>• Comments note the importance of studying and addressing differences in race, ethnicity, and gender. |

| Use and Usability (n=21) | Strengths | Limitations |
|---|---|---|
| **No Consensus**<br>24% Met<br>71% Not Met, but Addressable<br>5% Not Met | • The measure is in use in the Outpatient Quality Reporting Program.<br>• The usability of this measure is well established by the availability for reporting after several years.<br>• Feedback has been solicited on the measure. The measure has a feedback mechanism with no unintended consequences identified. | • Performance scores continue to show room for improvement but have remained largely stable from 2015-2021. It would be interesting to know how many sites have increased over the 6 years or remain at a stable level or have declined.<br>• The reasons for the lack of improvement are not clearly articulated. There's a need for the measure steward to address barriers to improvement.<br>• Question of the measure's utility for low-volume providers and whether the resources invested in it are positively impacting quality.<br>• There has been no substantial feedback or indications of unexpected findings. The only recommended intervention |

| Use and Usability (n=21) | Strengths | Limitations |
|---|---|---|
| | | is training providers, but there's a need for potential Quality Improvement (QI) mechanisms such as providing performance reports to providers. |

## CBE #4045 - Waveform Capnography in Ventilated Patients: Percent of patient transport contacts with advanced airways in whom continuous waveform capnography was used

**Number of Committee Reviews:** 21

| Importance (n=21) | Strengths | Limitations |
|---|---|---|
| **No Consensus**<br><br>29% Met<br><br>71% Not Met, but Addressable<br><br>0% Not Met | • **Waveform Capnography:** Standard of care for safe airway positioning and beneficial for patients.<br>• Support for using it in transport includes empirical studies and three consensus statements.<br>• The developer demonstrated the importance of using waveform capnography to recognize dislodgment of airways and to prevent serious complications to patients. | • **Unclear Specifications:** Unclear numerator qualifications<br>• **Lack of/Limited Evidence:** There is no systematic review of waveform capnography in transport.<br>• There is not a clear relationship in the logic model or the evidence for importance for how “higher score” translates into better quality.<br>• **Performance Gap:** Measure performance since 2014 ranges from 89.2% to 95.6% (87.5-94.1% among pediatric patients and 94-97.8% among adults) and may have limited room for improvement for some groups.<br>• **Meaningfulness to Patients:** Meaningfulness to patients has not been established. |

| Feasibility (n=21) | Strengths | Limitations |
|---|---|---|
| **No Consensus**<br><br>29% Met | • **Data Availability and Collection:** The measure is already in routine use by a substantial number of GAMUT users, which is compelling evidence that data is available and able to be captured. | • **Data Availability and Collection:** The most significant challenge is data capture. Data for this measure are not routinely generated from electronic sources but must be manually abstracted by hand from the EHR.<br>• There’s also a concern about manual abstraction and not using the available EHR or mobile capabilities to |

| Feasibility (n=21) | Strengths | Limitations |
|---|---|---|
| 67% Not Met, but Addressable<br><br>5% Not Met | • Required data are routinely generated and used during care and are available in EHRs or other electronic sources.<br><br>• **Feasibility:** Feasibility is demonstrated through normative evaluations from the consensus developers as well as the quality evaluators. | interface/upload to be able to pull the data necessary for this reporting measure.<br><br>• **Feasibility:** There's no assessment if the tools for capnography are easily available to transport service providers and if they are sufficiently easy to implement. The developers did not describe the process and how feasibility was done. It's unclear why the sites that have implemented this measure have not been able to develop documentation in the EHR or use mobile capabilities to have the data available to limit manual abstraction. |

| Scientific Acceptability (n=21) | Strengths | Limitations |
|---|---|---|
| **Consensus**<br><br>10% Met<br><br>86% Not Met, but Addressable<br><br>5% Not Met | • **Measure Definition:** The measure is well-defined and precisely specified. Details for measure calculation were clear and thorough.<br><br>• **Reliability:** The reliability of the measure appears satisfactory. The results suggest the measure is reliable with 100% agreement for one of the data elements and a Kappa value of 0.79 indicating substantial agreement for the other data element.<br><br>• **Helpful for Patients and Providers:** The measure appears to be valid and helpful for patients and healthcare providers. | • **Limited Testing:** The reliability was only tested at three sites. More facilities could be included in reliability testing.<br><br>• Not all data elements required for measure calculation were tested.<br><br>• **Denominator Exclusions:** There are concerns that the denominator exclusions are for the database manager to decide, which lacks standardization.<br><br>• **Appropriate Use of Capnography:** The interrater reliability should look at whether the reported data and |

| Scientific Acceptability (n=21) | Strengths | Limitations |
|---|---|---|
| | | the audited data were in agreement, not whether the use of capnography was "appropriate".<br>• **Validity:** There are concerns about the interpretation of the survey results for face-value validity, the details on face validity assessment, and the total number of randomly selected participants.<br>• There are also concerns about bias as face validity was done with current participants of the program, not an independent group of experts.<br>• It also does not discuss the reasons behind a high percentage of respondents do not agree. |

| Equity (n=21) | Strengths | Limitations |
|---|---|---|
| **No Consensus**<br>14% Met<br>14% Not Met, but Addressable<br>71% Not Met | • **Inclusivity:** The measure includes all patients who are transported with ventilators.<br>• **Standard of Care:** There should be no variation in equity since this is a standard of care. | • **Equity Considerations:** Lack of equity considerations in the measure. It suggests that demographic data could be used to assess inequities. |

| Use and Usability (n=21) | Strengths | Limitations |
|---|---|---|
| **No Consensus**<br><br>62% Met<br><br>33% Not Met, but Addressable<br><br>5% Not Met | • **Proven Usability:** The measure is already in use in many organizations, proving its usability.<br><br>• **Standard of Care:** The measure is a standard of care in all settings.<br><br>• **Existing Database**: The measure is currently in use within the GAMUT database since 2014 for internal QI, and QI with external benchmarking. | • **Lack of Patient and Family Input:** The absence of patient and family input might impact the understanding of improving this measure and maintaining measure improvements.<br><br>• **Lack of Details:** There is no clear description of planned uses within usability. In other sections, usability and use seem to be focused on consensus statements.<br><br>• **Limited Information:** This measure has been in use for several years. However, there is very little information regarding the measure being used in the past. |


Partnership for
Quality Measurement
Powered by Battelle