National Consensus Development and Strategic Planning for Health Care Quality Measurement

Guidance for the Closing Care Gaps Domain for the Endorsement and Maintenance (E&M) of Clinical Quality Measures

August 2025

Prepared by:
Battelle
505 King Avenue, Columbus, Ohio 43201

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. *Restricted:* Use, duplication, or disclosure is subject to the restrictions as stated in Contract Number 75FCMC23C0010 between the Government and Battelle.

Table of Contents

List of Tables

List of Figures

Introduction

Health care quality measurement is essential for evaluating and enhancing the performance of health care providers and systems in the U.S., which helps ensure that patients receive safe, effective, and patient-centered care. Despite its importance, traditional quality measurement has often failed to address the differences in health outcomes among different subpopulations.

The Partnership for Quality Measurement (PQM), formed by Battelle, a consensus-based entity (CBE) certified by the Centers for Medicare & Medicaid Services (CMS), plays a crucial role in shaping the health care quality measurement landscape. The PQM endorsement and maintenance (E&M) process reviews measures to ensure they are evidence based, scientifically sound, and both safe and effective (i.e., use of the measure will increase the likelihood of desired health outcomes, will not increase the likelihood of unintended and adverse health outcomes, and is consistent with current professional knowledge).

While the Closing Care Gaps domain remains voluntary through the Fall 2026 E&M cycle, this domain encourages the collection of data across different subpopulations, allowing accountable entities to assess differences, explore underlying causes, and promote accountability among health care providers and systems, thereby building trust and credibility within the patient community (i.e., patients, caregivers, and advocates).

Objectives and Purpose

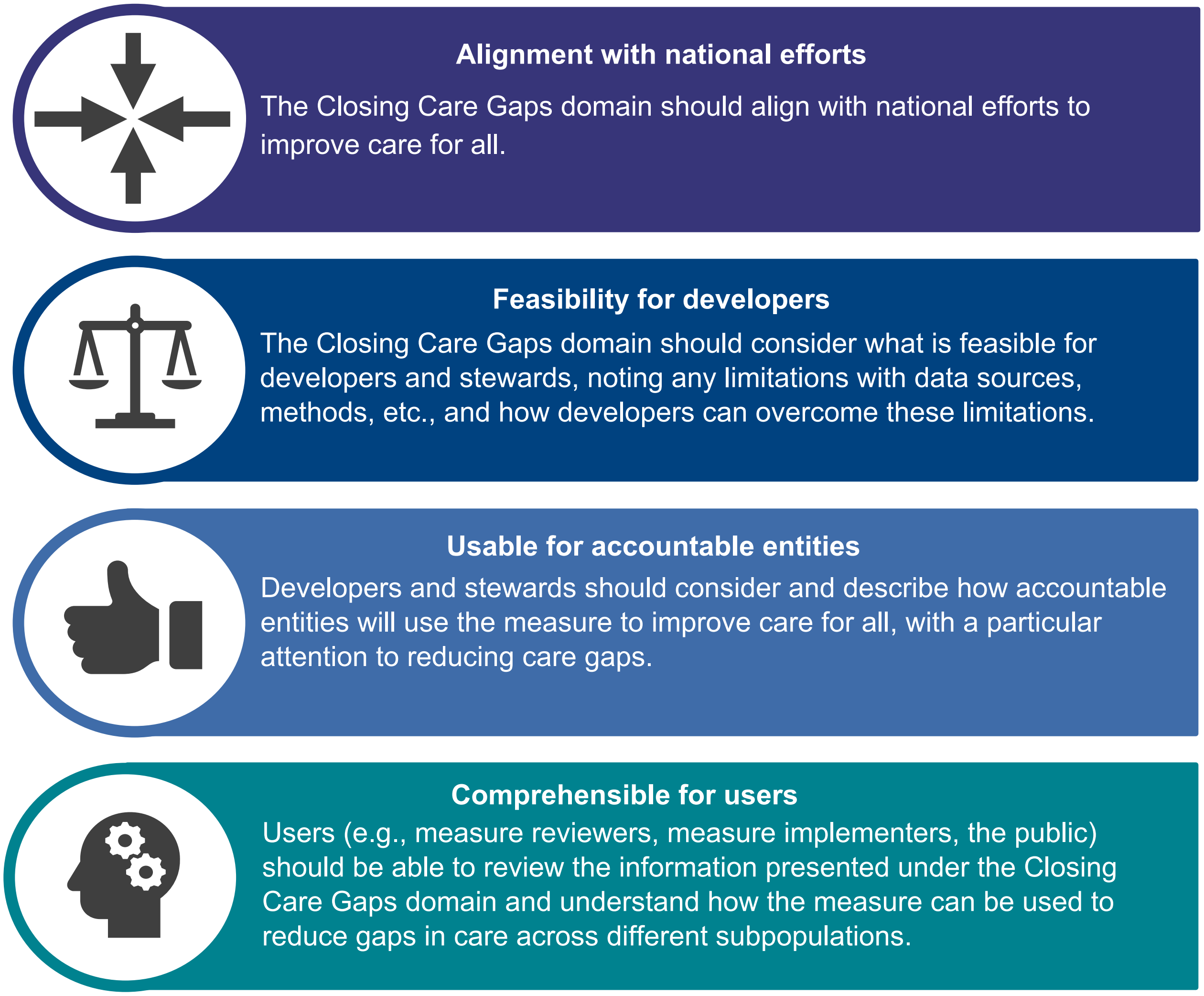
This document provides guidance to measure developers so they can accomplish the following:

- Describe the intent and key elements of the Closing Care Gaps domain and the expectations for initial and maintenance endorsement with respect to the PQM Measure Evaluation Rubric; and
- Identify examples of common methods, pitfalls, and mitigation approaches for supporting the claim that the measure contributes to reducing variations in health care and health outcomes.

Version 1.0 | August 2025 | *The analyses upon which this publication (or document) 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. Restricted:* Use, duplication, or disclosure is subject to the restrictions as stated in Contract Number 75FCMC23C0010 between the Government and Battelle.

Guiding Principles

PQM’s guiding principles for improving care for all are as follows:



Alignment with national efforts

The Closing Care Gaps domain should align with national efforts to improve care for all.

Feasibility for developers

The Closing Care Gaps domain should consider what is feasible for developers and stewards, noting any limitations with data sources, methods, etc., and how developers can overcome these limitations.

Usable for accountable entities

Developers and stewards should consider and describe how accountable entities will use the measure to improve care for all, with a particular attention to reducing care gaps.

Comprehensible for users

Users (e.g., measure reviewers, measure implementers, the public) should be able to review the information presented under the Closing Care Gaps domain and understand how the measure can be used to reduce gaps in care across different subpopulations.

Figure 1: PQM Guiding Principles for Closing Care Gaps

Audience

This resource guide supports measure developers by describing the expectations of the Closing Care Gaps domain within the PQM Measure Evaluation Rubric for CBE endorsement of clinical quality and cost/resource use measures. While examining and developing measures, measure developers should seek to answer the following questions:

- How does the measure contribute to efforts to address differences in health care delivery and health outcomes?
- How can accountable entities use the measure to improve care for all and reduce differences in care delivery and outcomes?

Closing Care Gaps Definition and Purpose

Purpose of Closing Gaps in Care

Closing Care Gaps

The process of identifying and addressing differences in health care access, quality, and outcomes among various subpopulations.

For the purposes of CBE endorsement, “closing care gaps” refers to the process of identifying and addressing differences in health care access, quality, and outcomes among various subpopulations.[1,2] These differences may be associated with factors such as age, sex, race, ethnicity, geography, language preference, income level, or other demographic characteristics that influence a person’s ability to receive timely, appropriate, and effective care.

By focusing on closing care gaps, clinical quality and/or cost/resource use measures can help ensure that care is consistently delivered across all populations, and that no group is systematically left behind in achieving positive health outcomes.

Operational Definitions

Socio-contextual: Social context includes the human interactions where one lives, works, learns, plays, and worships.[3]

Subpopulation: A specific group of individuals with common patient characteristics (e.g., sex age, rural or urban residence, risk factors) that is the target of an intervention or a policy recommendation.

Technical expert panel (TEP): A group of interested parties with specialized knowledge and experience relevant to the specific health care quality measure being developed. This panel typically includes a mix of clinicians, statisticians, quality improvement specialists, methodologists, pertinent measure developers, and consumers.

Variable: A data element used in analysis, including the final risk model or stratification plan. Variables can include characteristics such as demographics, geographic location, and others important for stratifying clinical quality measure results by subpopulations.

---

[1] Centers for Medicare & Medicaid Services. (2025, May 12). Preserving Medicaid Funding for Vulnerable Populations – Closing a Health Care-Related Tax Loophole Proposed Rule. Retrieved from https://www.cms.gov/newsroom/fact-sheets/preserving-medicaid-funding-vulnerable-populations-closing-health-care-related-tax-loophole-proposed.

[2] Office of the Assistant Secretary for Planning and Evaluation (ASPE), U.S. Department of Health and Human Services. (2025, March). Health Care and Child Care Costs Contribute to the Unsustainable and Growing Cost of Raising a Family in America. Retrieved from https://aspe.hhs.gov/reports/health-care-child-care-costs.

[3] DHHS Office of Disease Prevention and Health Promotion. Healthy People 2030: Social and Community Context. Retrieved October 8, 2024 from https://health.gov/healthypeople/objectives-and-data/browse-objectives/social-and-community-context.

Elements of the Closing Care Gaps Domain

This section provides an in-depth look at the key elements of the Closing Care Gaps domain, including expectations for initial vs. maintenance endorsement, pitfalls, and mitigation approaches.

Consult External Sources

To address the Closing Care Gaps domain, measure developers should consider consulting with external sources such as patients and caregivers, population health experts, and accountable entities. This consultation helps ensure that the measure is relevant and actionable.

The Closing Care Gaps domain is comprised of four elements (Figure 2). Each has a different expectation depending on whether the measure is being considered for initial or maintenance endorsement. Overall, the Closing Care Gaps domain attempts to identify the extent to which the measure can distinguish differences in care for certain patient subpopulations and the extent to which accountable entities can use the measure to improve care for all. Developers should adopt a comprehensive approach that includes empirical analysis, theoretical reasoning, experiential insights, and logical deduction, which ensures a more robust assessment of care gaps.

To address the Closing Care Gaps domain, measure developers should consider consulting with external sources such as patients and caregivers, population health experts, and accountable entities. This consultation helps ensure that the measure is relevant and actionable. Appendix A contains additional resources to assist with developing responses for the Closing Care Gaps domain.

Element 1 – Evidence of Known Differences

Identify and describe known and existing variations in health care and health outcomes related to the measure focus area. Include factors such as demographic characteristics and groups that have historically experienced barriers to accessing health care. Describe any challenges and how the measure contributes to efforts to improve health care delivery and outcomes for these identified groups.

Element 2 – Testing Methods

Provide a description of your methodology and approach to empirical testing of differences in performance scores across identified groups (e.g., demographic, geographic variables).

Element 3 – Results and Interpretation

Provide the results and an interpretation of those results, including explanations for any variations in performance scores across different groups. Discuss how these results relate to existing evidence (Element 1), any limitations found in the results, and the potential impact of these variations on the identified groups.

Element 4 – Anticipated Impact

Describe or provide evidence indicating how accountable entities can utilize these results to close gaps in health care delivery and outcomes for the identified groups.

Figure 2: Four Elements of the Closing Care Gaps Domain

Initial and Maintenance Endorsement Expectations

Though submitting input for this domain is voluntary, for developers completing the Closing Care Gaps Domain, there are specific expectations based on endorsement status (i.e., initial endorsement or maintenance endorsement). For initial endorsement, measure developers should address Element 1 (Evidence of Known Differences). While Elements 2 (Testing Methods), 3 (Results and Interpretation), and 4 (Anticipated Impact) are optional, developers are encouraged to address them during the initial endorsement process if possible (Figure 3).

For maintenance endorsement, measure developers are encouraged to complete all elements of the Closing Care Gaps domain, if information is available. For subsequent maintenance reviews, measure developers should consider the following questions and provide supportive evidence:

- What progress, if any, has been made to address previously identified subpopulation differences with respect to the measure focus area?
- What changes in the performance results across identified subpopulations have occurred?
- What new identified subpopulation differences have arisen since the measure was last endorsed?

Closing Care Gaps Domain Examples

To further support measure developers, this guidance provides two Closing Care Gaps domain examples—one within the body of this guidance and the second within Appendix B—from measures submitted to Battelle for endorsement. We adapted the responses to the Closing Care Gaps domain for both measures to further align with the expectations identified in this guidance document.

The first measure, CBE #4220 Breast Cancer Screening Recall Rates, is a process measure that uses claims data to calculate the percentage of beneficiaries with mammography or digital breast tomosynthesis (DBT) screening studies that are followed by a diagnostic mammography, DBT, ultrasound, or magnetic resonance imaging (MRI) of the breast in an outpatient or office setting on the same day or within 45 days. The second measure, CBE #0133 In-Hospital Risk Standardized Mortality for Percutaneous Coronary Intervention (PCI), is an outcome measure that uses data from a registry to identify rate of hospital-level risk-standardized mortality in adult patients without cardiogenic shock or cardiac arrest undergoing PCI.

These measures were selected because each contained responses that aligned with what is outlined in this guidance. The selection of these measures as examples had no effect on the measures' endorsement decisions.

Element 1 – Evidence of Known Differences*

Initial Endorsement

- Review literature and other evidence (e.g., discussions with TEPs, advisory groups, internal analyses) to identify subpopulation differences with respect to the measure focus.

Maintenance Endorsement

- Engage in a similar process to the initial endorsement phase, evaluating for any new subpopulation differences and/or improvement in previously identified subpopulations.

Element 2 – Testing Methods

Initial Endorsement (optional)

- Explain methodology and approach to empirical testing differences in performance scores across identified subpopulations.
- Indicate which subpopulations will be evaluated, noting any data limitations.

Maintenance Endorsement (if data are available)

- Determine if the previously used methodology should be used with more recent data to update the results or if a new method should be used. If using a new method, explain why.

Element 3 – Results and Interpretation

Initial Endorsement (optional)

- Report the performance results for the subpopulations identified in Element 1 using the methodology described in Element 2.
- Interpret findings with emphasis placed on the impact to identified subpopulations.
- Discuss consistency of findings with Element 1's evidence and literature base.

Maintenance Endorsement (if data are available)

- Re-evaluate updated data.
- Determine if the results have any differences compared to the prior submission and include a rationale for the differences.

Element 4 – Anticipated Impact

Initial Endorsement (optional)

- Based on findings from Element 3, provide evidence-based recommendations of approaches providers can take to address identified subpopulation differences.

Maintenance Endorsement (if information is available)

- Examine previously submitted recommendations and update accordingly.

Figure 3: Elements of the Closing Care Gaps Domain – Initial vs. Maintenance

Element 1 – Evidence of Known Differences

Identify and describe from existing literature, internal analyses, etc. known and existing differences in health care and health outcomes across patient subpopulations (e.g., demographics, geography, insurance type) related to the measure focus area. Discuss how the findings relate to the measure and how the measure may contribute to efforts to improve health care delivery and outcomes for these identified groups.

To the extent feasible, measure developers may consider variables highlighted in Table 1. Variables identified in the table align with national efforts to address care gaps and are generally accessible for analysis.[1] Developers may consider these variables as potential subpopulations for analysis in Elements 2-4. This consideration does not mean that developers are required to use these variables if evidence suggests otherwise (Table 2) or if there are data limitations to conducting a stratification analysis (see Element 2). This list is also not meant to limit measure developers from exploring other variables relevant to the specific measure focus. If additional variables or subpopulations are to be considered, these should be noted for Element 1. Additionally, developers should describe any current or past efforts to address these differences, including if care gaps remain and how the measure will contribute to these efforts.

Table 1: Example Subpopulation Variables

Variable*	Example of Variable**
Age	• Age (in years)[o] • Date of birth[o]
Race	• Beneficiary race code (Unknown, White, Black, Other, Asian, Hispanic, North American Native)[o]
Ethnicity	• Ethnicity code (Not of Hispanic or Latino/a or Spanish origin; Mexican, Mexican American Chicano/a; Puerto Rican; Cuban; Another Hispanic, Latino, or Spanish origin; Hispanic or Latino, Subcategory Unknown)[o]
Sex	• Sex (Male, Female)[o]
Insurance Coverage	• Payer type code (Medicare only, Medicaid only, dually eligible for Medicare and Medicaid, missing)[o]
Language	• Preferred language (Arabic, Armenian, Chinese, English, French, German, Greek, Gujarati, Hebrew, Hindi, Italian, Japanese, Khmer, Korean, Other, Persian, Polish, Portuguese, Russian, Serbo-Croatian, Spanish, Tagalog, Thai, Urdu, Vietnamese, Yiddish, Missing)[o] • English language proficiency code (Very Well, Well, Not Well, No Spoken Proficiency, Missing)[o]
An Indicator of Urbanicity/Rurality	• ZIP code, ZIP code tabulation areas (ZCTAs), population density, adequate number of providers, designation as a medically underserved area[^]

Variable*	Example of Variable**
An Indicator of Disability (e.g., frailty)	- Frailty, results from standardized assessments (e.g., Activities of Daily Living [ADL] and Instrumental Activities of Daily Living [IADL] scales)+
An Indicator of Socioeconomic (SES) Status (e.g., SES indices)	- Income based on federal poverty guidelines, education, occupation, income, housing security, percent uninsured/Medicaid°

** As much as possible, measure developers should attempt to align definitions with those provided on a national level (e.g., U.S. census).*
*** Examples listed can be identified within various data sources, such as claims, electronic health records, the U.S. census, etc. For more information on these data sources, please see* Appendix D and the following:
° *"Search Data Variables." ResDAC, University of Minnesota,* https://resdac.org/search-data-variables
^ *"Rural-Urban Commuting Area Codes." Economic Research Service, U.S. Department of Agriculture,* https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.
+*Developing and Assessing the Validity of Claims-based Indicators of Frailty & Functional Disabilities in Electronic Health Records. Content last reviewed September 2022. Agency for Healthcare Research and Quality, Rockville, MD.* https://www.ahrq.gov/learning-health-systems/frailty-ehr.html.

Maintenance Endorsement

Measure developers should evaluate any updated literature, internal empirical analyses, and/or qualitative data to determine if any changes in previously identified differences have occurred since the measure's last endorsement review, including whether new subpopulation differences have emerged. Use previously reported results as benchmarks to assess progress. If no change has occurred, developers should provide a rationale as to why.

Table 2: Element 1 Common Challenges and Mitigation Approaches

Challenge	Mitigation Approach
How many subpopulations should be considered?	While there is no requirement for a specific number of subpopulations, Table 1 provides a list of variables developers may consider. Developers should explore subpopulations that are relevant to the specific measure focus, including those that have known and existing differences in care.
There is no evidence of differences, or there are no data available.	Measure developers should explain how they determined that there was no evidence of differences by summarizing findings from literature reviews, internal empirical analyses, and/or qualitative data. Note gaps in evidence or inconsistent evidence. Measure developers should explain their review process and sources of evidence.
Intrinsic bias of available data exists.	Measure developers should actively seek evidence that includes diverse populations and acknowledge evidence limitations (e.g., lack of available evidence for certain subpopulations). Consider engaging a TEP or representatives from affected communities to gain further insight.

Element 1 Example

The measure CBE #4220 Breast Cancer Screening Recall Rates will be used throughout the main body of this document as an example. The example was developed for guidance purposes, does not reflect the actual work of its measure developer/steward, and had no impact on the measure's endorsement decision. Appendix B contains an additional example.

Measure Description: *The Breast Cancer Screening Recall Rates measure calculates the percentage of beneficiaries with mammography or digital breast tomosynthesis (DBT) screening studies that are followed by a diagnostic mammography, DBT, ultrasound, or magnetic resonance imaging (MRI) of the breast in an outpatient or office setting within 45 days.*

Initial Endorsement

From the perspective of both clinical quality and efficiency, potentially negative consequences exist if the mammography and DBT recall rate is either too high or too low. A high cumulative dose of low-energy radiation can be a consequence of too many false-positive mammography and DBT follow-up studies. Radiation received from mammography or DBT may induce more cancers in younger people or those carrying deleterious gene mutations, such as BRCA-1 and BRCA-2.[4]

Professional societies and guidelines provide inconsistent suggestions on the appropriate recall rates to establish for breast cancer screening. The American College of Radiology (ACR) recommends a target recall rate for mammography screening between 5% and 12%.[5] European research, via the International Agency for Research on Cancer, sets a target recall rate of 5%.

Significant research has focused on breast cancer screening and which specific subpopulations experience lower rates of screening. In 2024, the U.S. Preventive Services Task Force (USPSTF) revised its breast cancer screening guidelines, now advising that all women start mammography screenings every 2 years beginning at age 40 and continuing through age 74. This change aims to address the higher mortality rates and earlier onset of aggressive breast cancer subtypes, such as triple-negative cancers, among Black women, despite their lower lifetime incidence of the disease.[6]

---

[4] Berrington de Gonzalez, A., Berg, C., Visvanathan, K., & Robson, M. (2009). Estimated risk of radiation-induced breast cancer from mammographic screening for young BRCA mutation carriers. *JNCI: Journal of the National Cancer Institute, 101*(3), 205–209. https://doi.org/10.1093/jnci/djn440

[5] D'Orsi, C. J., Sickles, E. A., Mendelson, E. B., Morris EA, et al. (2013). ACR BI-RADS® atlas, breast imaging reporting and data system. Reston, VA: American College of Radiology.

[6] United States Preventive Services Taskforce. Breast cancer: screening. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/breast-cancer-screening#bcei-recommendation-title-area. Published April 30, 2024.

In addition to the clinical quality and efficiency concerns associated with mammography and DBT recall rates, variations in these rates by race, age, and dual eligibility further complicate the effectiveness of breast cancer screening programs. Studies have shown that white women and Hispanic women tend to have higher recall rates for additional imaging following their initial mammography compared to Black women. This care gap can be attributed to various factors, including differences in breast density, access to health care, and socioeconomic factors.[7] Younger women, typically those under 50, experience higher recall rates due to denser breast tissue, which can obscure mammographic images and lead to more frequent callbacks for additional testing.[8] Additionally, a recent study on the performance of an artificial intelligence (AI) algorithm in interpreting negative screening DBT examinations found that false-positive case scores were significantly more likely in Black women and older patients, indicating that AI tools might perpetuate existing differences.[9]

Dual eligibility, representing individuals who qualify for both Medicare and Medicaid, often indicates a lower SES and is associated with differences in recall rates. These patients may face multiple additional barriers to effective breast cancer screening, including limited access to high-quality mammography services and follow-up care. Research indicates that women with dual eligibility are less likely to receive timely follow-up after abnormal mammograms compared to their non-dual-eligible counterparts, suggesting that socioeconomic factors compound the challenges faced by these women.[10]

To ensure that breast cancer screening is safe and effective, addressing these variations is crucial. This measure aligns with national efforts to reduce differences in populations that are typically screened at a lower rate and recalled at different rates.[11] This measure can facilitate more standardization of recall rates, monitor AI algorithm performance and the potential differences it may exacerbate, and increase provider awareness of where differences exist within their patient populations. These efforts will not only improve radiologist training to recognize and mitigate biases in image interpretation but also enhance patient communication

---

[7] Lee, C.S., Goldman, L., Grimm, L.J. *et al.* Screening mammographic performance by race and age in the National Mammography Database: 29,479,665 screening mammograms from 13,181,241 women. *Breast Cancer Res Treat* 203, 599–612 (2024). https://doi.org/10.1007/s10549-023-07124-6

[8] Smith-Bindman, R., Lipson, J., Marcus, R., et al. (2012). Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Archives of Internal Medicine, 172(22), 1813-1820. DOI: 10.1001/archinternmed.2013.11963.

[9] Nguyen, Derek L., et al. “Patient Characteristics Impact Performance of AI Algorithm in Interpreting Negative Screening Digital Breast Tomosynthesis Studies.” Radiology, vol. 311, no. 2, 2024.

[10] Walker, J., Techakehakij, W., Smith, D. R. (2018). Delayed follow-up of abnormal breast cancer screening among women with dual Medicaid and Medicare eligibility. Journal of Women's Health, 27(7), 904-913. DOI: 10.1089/jwh.2017.6575

[11] Evidence-Based Cancer Control Programs. National Cancer Institute. The witness project. https://ebccp.cancercontrol.cancer.gov/programDetails.do?programId=270521. Updated July 31, 2020.

and follow-up procedures. Together, these initiatives ensure that all women, regardless of race, age, or SES, receive appropriate and timely care following mammography screenings.

Maintenance Endorsement

For measure maintenance, the Element 1 information from prior endorsement may serve as a starting point. This section includes additional information that should be provided.

In reviewing current literature for maintenance endorsement, there has been little to no movement in a positive or negative direction in relation to the prior identified differences. The performance score stratification results for this measure show slight improvement when compared to the last endorsement review, but the differences in race, age, and dual eligibility were still statistically significant (see Table E2 under Element 3). This is likely due to a continued care access issue, as Black women are more likely than white women to experience longer delays to definitive diagnosis and treatment and receive less guideline-concordant care.[12] The statistically significant difference in regards to age is reflective of the literature, which indicates that younger women have higher recall rates. The performance score data also confirms the dual-eligibility differences in the literature. All this information indicates that the differences highlighted in the literature persist, meaning the measure's purpose of encouraging providers to undertake internal quality improvement efforts to address these differences is still important.

Element 2 – Testing Methods

If data are available, provide a description of your methodology and approach to empirical testing of differences in performance scores across identified groups (e.g., demographic, geographic variables).

Measure developers should identify which variables from Element 1 they intend to analyze. This includes considering limitations of available data sources that prohibit the ability for further analysis. In addition, within Section 1 (Measure Specifications) of the PQM Full Measure Submission Form, developers should document and disclose the data sources utilized, including the dates of data collection, any data cleaning and manipulation processes undertaken, and the quality of the data (for example, whether the data undergo periodic audits).

Measure developers should also describe the methodology used to empirically test differences in performance scores across the subpopulations identified from Element 1. This description should include a rationale for using the method and a detailed outline of the steps involved in the methodology. When feasible, present a distribution of the measure scores and sample sizes for each subpopulation. There may be instances where a statistically significant variable is identified as endemic to a population and therefore not actionable. To address this, measure

---

[12] Brawley OW, Lansey DG. Disparities in breast cancer outcomes and how to resolve them. Hematol Oncol Clin North Am. 2023;37(1):1-15. doi:10.1016/j.hoc.2022.08.002

developers could consider involving the communities being studied by utilizing a population index (see resource in Appendix A; Table 3).

Results from the analyses may inform other areas of the measure submission, such as risk adjustment or stratification, and should align across the submission. However, the purpose of the Closing Care Gaps domain is not to justify the inclusion or exclusion of variables in risk adjustment but to assess whether the measure can identify differences in care for certain patient populations. If a variable identified in Element 1 is included in the risk model, then the variable should be taken out of the risk model; the risk model is then run, and the measure performance is then stratified by the variable of interest. Measure developers may refer to the stratification approach previously discussed in their submission, drawing attention to any additional variables identified in Element 1.

Selecting a Method

When selecting a method to empirically test the differences in performance scores across identified variables, measure developers should consider the type of data (e.g., continuous, discrete/categorical, proportions) and how those data are distributed. For example, race and ethnicity are discrete/categorical data, and the chi-square test would be best utilized for the empirical analysis. Appendix C provides a flowchart of common methods for empirical tests. In addition, it is best practice to adhere to a minimum confidence interval of 95%, which means that for each stratum, the calculated interval (e.g., for the mean, proportion) is expected to contain the true population parameter for that stratum 95% of the time. This approach helps in making reliable inferences about each subpopulation, considering their unique characteristics. In addition, stratification segments should not be so small that they are no longer statistically significant. Some limitations for this work may include insufficient data points to stratify the measure properly. Lastly, the Closing Care Gaps domain adheres to CMS's cell-suppression policy, which stipulates that no cell (e.g., admissions, discharges, patients, services) containing a value of less than 11 can be reported directly.

Data Limitations

Data availability may vary by data source and could be an added burden to the accountable entity. If a variable identified in Element 1 is not readily available within a data source, measure developers should note this limitation in their response to Element 2. Developers are encouraged to engage with accountable entities, as feasible, to better understand any resource challenges (e.g., data availability, staffing constraints) and are encouraged to provide guidance on collecting data for identified variables to mitigate any potential burden. This is especially relevant for non-claims data sources (e.g., electronic health record data, registry data; see Table 3).

Maintenance Endorsement

During maintenance endorsement, the developer should re-examine the methodology and approach used to empirically test differences in performance scores across identified variables to determine if they are still appropriate given any additional variables identified. If a different approach is needed, developers should describe the method and rationale for why the approach is being used. Variables evaluated during the measure's last endorsement review, as well as any newly identified variables, should be empirically assessed.

Table 3: Element 2 Common Challenges and Mitigation Approaches

Challenge	Mitigation Approach
Variables identified as significant are viewed as endemic to populations surveyed.	Explain the impact of variables viewed as endemic. Measure developers should consider involving the communities being studied by utilizing a population index to increase understanding in various subpopulations. (See resources in Appendix A.)
There is a lack of available data to stratify performance scores by identified subpopulations.	Consider engaging with groups, such as a TEP or other accountable entities, to address gaps in data. This may include developing data-collection guidance and resources. Collaborate with other health care providers, research organizations, or community groups to access additional data. Leverage public databases or datasets that might provide relevant information for stratification. If related variables are available, consider using them as proxies for the missing data. For example, if specific socioeconomic data are missing, the developer might use geographical information as a proxy and assume certain socioeconomic characteristics based on location. Apply statistical methods to estimate the missing values based on the patterns and relationships observed in the available data. Techniques such as multiple imputation or regression imputation can be used, but acknowledging the potential for bias is important. Provided that the data limitations are thoroughly evaluated, it is acceptable if stratification by certain subpopulations is not feasible. By engaging in this process, developers help ensure their analysis remains transparent and well-informed despite the constraints. Potential data sources are included in Appendix D.

Element 2 Example

Initial Endorsement

Encouraged for initial endorsement.

Maintenance Endorsement

To empirically test differences in performance scores across identified subpopulations:

1. We first defined and identified the subpopulations to be analyzed in Element 1, which included patient sex, racial or ethnic identity, age, and dual-eligibility status.
2. We ensured that the data used for analysis came only from facilities that met the minimum case count requirements for public reporting. This step was crucial to maintain the reliability and validity of the analysis.
3. We then employed chi-square tests to assess the differences in measure performance across the defined subpopulations. The chi-square test was particularly useful in

determining if the observed differences in performance were significant, allowing us to identify whether the variations were due to chance or indicative of genuine differences.

4. We analyzed the chi-square values and associated probabilities to determine whether any observed variations in performance across subpopulations were statistically significant or due to chance. This helped identify specific areas where differences existed and where targeted interventions could be implemented.

For further maintenance reviews, the Element 2 information from prior endorsement may serve as a starting point. This section includes additional information that should be provided.

Our approach to stratifying performance scores to empirically test for differences across subpopulations is consistent with our approach used during the measure's last endorsement review. Chi-square tests have been consistently employed to assess the differences in the recall rates and to highlight any differences.

Element 3 – Results and Interpretation

If data are available, provide the results and an interpretation of those results, including explanations for any variations in performance scores across different groups. Discuss how these results relate to existing evidence (Element 1), any limitations found in the results, and the potential impact of these variations on the identified groups.

Measure developers should present and interpret their results—describing any differences found in the data, including those that are statistically significant or not (i.e., whether the observed difference is due to chance). Developers should also consider the clinical and community significance of the results, which refers to the practical importance of any differences identified and the relevance of the findings to the identified subpopulations, respectively.

The results interpretation should include a discussion of any positive or negative results for the variables analyzed. Measure developers should consider differences previously identified from the literature, internal empirical analyses, and/or prior endorsement submissions (as outlined in Element 1) and integrate these findings into their interpretation of current results. For instance, do current results indicate a narrowing or widening of performance gaps compared to what has been historically observed or anticipated? If a previously identified difference for a specific subpopulation widened, this deviation should be clearly reported along with rationales as to why. For example, if prior literature shows Black women have reduced breast cancer screening recall rates, but the stratification results do not show any significant differences, this could be due to a limited data set that may not be generalizable. So, a larger, more generalizable testing sample may show similar differences to what was identified in Element 1.

Developers should acknowledge and clearly document common limitations along with the results interpretation. Some common limitations may include, but are not limited to, sample size variability, data-collection variability, or confounding variables (Table 4). Smaller subpopulations may lead to smaller sample sizes, which can affect the reliability of the results. For example, performance measures for a rare disease might be based on a small number of cases, leading to greater variability in quality metrics. To mitigate this, developers and stewards may consider increasing the sample size to ensure more stable and reliable estimates. Increasing the sample size might involve extending the data-collection period or combining data from multiple sources.

Differences in data-collection methods among accountable entities can lead to inconsistencies. Developers should note any potential biases introduced by varied data-collection methods or interpretative processes used across different health care entities. To mitigate this limitation, developers may develop measure-implementation guidance that includes standardized protocols for data collection amongst accountable entities to minimize variability.

Lastly, it is essential to consider that unmeasured or unknown confounding variables, such as other socioeconomic factors or comorbidities, might be present. Developers should discuss how such factors could have impacted the findings. For instance, if a quality measure shows poorer outcomes for a specific racial group, consider whether differences in access to care or other underlying health care differences might have influenced these results. For outcome and cost/resource use measures, developers should consider comprehensive risk-adjustment models that include a wide range of demographic, socioeconomic, and clinical variables, which may control for potential confounders.

When documenting limitations, developers should provide specific examples to illustrate how certain factors might have influenced the results. This transparency helps users of the measures understand the context and potential biases affecting the data.

Maintenance Endorsement

In addition to the requirements noted above for Element 3, measure developers should continuously assess any improvements in the results and provide supporting rationales for these changes.

Table 4: Element 3 Common Challenges and Mitigation Approaches

Challenge	Mitigation Approach
A lack of statistically significant results exists.	All limitations should be clearly stated in Element 3. Consider other types of significance when interpreting results: • Statistical significance: The real and relevant deviations in data. • Clinical significance: The effectiveness of the data. • Community significance: Relevance to identified subpopulations.
Smaller subpopulations may lead to smaller sample sizes.	Developers may consider increasing the sample size to ensure more stable and reliable estimates. Increasing the sample size might involve extending the data-collection period or combining data from multiple sources.
There is variability present in data collection.	Note any discrepancies. Developers may develop measure implementation guidance that includes standardized protocols for data collection amongst accountable entities to minimize variability.
The presence of confounding variables, known or unknown, should be	Developers should articulate whether other factors may have influenced the results. For outcome and cost/resource use measures, developers should

Challenge	Mitigation Approach
considered.	explore whether comprehensive risk adjustment models that include a wide range of demographic, socioeconomic, and clinical variables can adequately control for potential confounders.

Element 3 Example

Initial Endorsement

Encouraged for initial endorsement.

Maintenance Endorsement

Data were collected in an elective manner as reported by participating facilities. As shown in Table E1, some potential social risk factors were examined to identify performance gaps. These factors include sex, racial or ethnic identity, age band, and dual-eligibility status. Statistically significant differences in performance have been identified, demonstrating an opportunity to improve care for all based on these risk factors.

Table E1: 2018 Performance Scores by Patient Sex, Racial or Ethnic Identity, Age Band, and Dual-Eligibility Status from Facilities Meeting Minimum Case Count Requirements**

Characteristic	Category	N	% of Denominator	Performance Score	Chi-Square Variable	Chi-Square Probability
Sex	Male	314	0.01%	20.1%	85.8878	<0.0001
	Female	3,287,982	99.9%	8.9%		
Racial or Ethnic Identity	Unknown	95,245	3.0%	11.0%	440.2351	<0.0001
	White (Non-Hispanic)	2,746,311	85.5%	9.0%		
	Black (Non-Hispanic)	224,645	7.0%	8.0%		
	Other	60,789	1.9%	9.2%		
	Asian or Pacific Islander	40,511	1.3%	10.3%		
	Hispanic or Latino	36,411	1.1%	9.0%		
	American Indian or Alaska Native	7,251	0.2%	6.0%		

Characteristic	Category	N	% of Denominator	Performance Score	Chi-Square Variable	Chi-Square Probability
Age Band	18-34	175	0.01%	5.0%	1290.789	<0.0001
	35-44	17,689	0.53%	16.5%		
	45-54	61,456	1.85%	14.9%		
	55-64	168,098	5.05%	10.0%		
	65-74	2,111,389	63.42%	9.1%		
	75-84	860,654	25.85%	9.1%		
	85+	109,876	3.30%	--		
Dual Eligibility	Medicare Only	3,023,395	92.9%	8.5%	5.578	<0.0001
	Dual Eligible	230,456	7.1%	8.4%		

**2018 data were generated for demonstration purposes for this guidance document and are not real values.

There may be some limitations in the interpretation of these data, as some categories make up a small percentage of the total for each characteristic. For example, only 0.01% of patients in the measure sample are male (as would be expected given the clinical scope of the measure), although the chi-square probability (<0.0001) indicates that the difference in performance (20.1% for males, 8.9% for females) is significant. Racial identity also provides a similar chi-square probability (<0.0001), with white patients making up most cases (85.5% of the total initial patient population, with a performance rate of 9.0%) followed by Black patients (7.0% of the initial patient population, with a performance rate of 8.0%). The next largest category is unknown race, comprising 3.0% of the initial patient population, with performance at 11.0%. Performance scores for patients of “other” race (9.2%), Asian or Pacific Islander (10.3%), and American Indian or Alaska Native (6.0%) show significant variation between race categories. Similarly, patients of Hispanic or Latino ethnicity (9.0%) also vary substantially from non-Hispanic or non-Latino populations.

Age band categories show consistent trends of lower scores as age increases, ranging from 16.5% for patients aged 35 to 44 years to 9.1% for those aged 75-84 years. Younger patients make up a small percentage of the overall testing population, with the categories including those aged 18 to 34 comprising about 0.01% of the initial patient population.

Finally, performance by dual eligibility was examined, with 92.9% of the initial patient population having only Medicare fee-for-service (FFS) coverage and the remaining 7.1% enrolled in both Medicare FFS and Medicaid (dually eligible). The difference in performance was slight—8.5% for Medicare only versus 8.4% for dually eligible—but significant at the 0.05 level (p<0.0001).

These findings are consistent with the information provided in Element 1, which highlighted differences between White and Black subpopulations and between populations with dual eligibility and populations that only have Medicare coverage. While there have been improvements, the rates for these subpopulations continue to be higher compared to other demographics and continue to have clinical and community significance.

For further maintenance reviews, the Element 3 information provided during prior endorsement may serve as a starting point. This section includes additional information that should be provided.

There have been slight changes in the performance scores from 2018 to 2023, but none of the population scores in the subpopulations demonstrated statistically significant changes at the 0.05 level (Table E2). The minute change in performance scores suggests that efforts to address differences in recall rates in regard to race, sex, age, insurance eligibility between 2018 and 2023 have not been enough to address the discrepancies in the data in a statistically significant way.

Table E2: Performance Scores by Patient Sex, Racial or Ethnic Identity, Age, and Dual-Eligibility (2018 vs. 2023)**

Characteristic	Category	N (2018)	N (2023)	% of Denominator (2018)	% of Denominator (2023)	Performance Scores (2018)	Performance Scores (2023)	Chi-Square Value (2018)	Chi-Square Value (2023)	Chi-Square Probability (2018)	Chi-Square Probability (2023)
Sex	Male	314	317	0.00%	0.0%	20.1%	24.3%	85.8878	86.2587	<0.0001	<0.0001
	Female	3,287,982	3,307,543	99.9%	99.9%	8.9%	9.2%				
Racial or Ethnic Identity	Unknown	95,245	67,684	3.0%	2.1%	11.0%	10.7%	440.2351	441.4312	<0.0001	<0.0001
	White (Non-Hispanic)	2,746,311	2,859,308	85.5%	86.4%	9.0%	9.2%				
	Black (Non-Hispanic)	224,645	250,620	7.0%	7.6%	8.0%	8.5%				
	Other	60,789	40,434	1.9%	1.2%	9.2%	9.8%				
	Asian or Pacific Islander	40,511	45,795	1.3%	1.4%	10.3%	10.0%				
	Hispanic or Latino	36,411	36,398	1.1%	1.1%	9.0%	9.4%				
	American Indian or Alaska Native	7,251	7,619	0.2%	0.2%	6.0%	6.7%				
Age Band	18-34	175	193	0.01%	0.0%	5.0%	6.7%	1290.789	1334.694	<0.0001	<0.0001
	35-44	17,689	18,271	0.53%	0.6%	16.5%	17.6%				
	45-54	61,456	62,644	1.85%	1.9%	14.9%	15.1%				
	55-64	168,098	162,033	5.05%	4.9%	10.0%	11.0%				
	65-74	2,111,389	2,078,294	63.42%	62.8%	9.1%	9.3%				

Characteristic	Category	N (2018)	N (2023)	% of Denominator (2018)	% of Denominator (2023)	Performance Scores (2018)	Performance Scores (2023)	Chi-Square Value (2018)	Chi-Square Value (2023)	Chi-Square Probability (2018)	Chi-Square Probability (2023)
	75-84	860,654	875,879	25.85%	26.5%	9.1%	9.3%				
	85+	109,876	110,546	3.30%	3.3%	--	8.8%				
Dual Eligibility	Medicare Only	3,023,395	3,062,303	92.92%	92.6%	8.5%	9.2%	5.578	5.787	<0.0001	0.0161
	Dual Eligible	230,456	245,514	7.08%	7.4%	8.4%	9.3%				

**2018 and 2023 data were generated for demonstration purposes for this guidance document and are not real values. 2018 values are taken from Table E1 in the Element 3 example.

Element 4 – Anticipated Impact

If information is available, discuss how accountable entities have utilized the measure to close gaps in health care delivery and outcomes for the identified groups.

Measure developers should provide a summary explaining how accountable entities can utilize stratified results to implement interventions aimed at reducing differences in care across identified subpopulations and to evaluate the interventions' effectiveness. This summary should consider the varying capacities and resources of different accountable entities, such as the differences between small community clinics and large health care systems (Table 5). It should also emphasize the potential for collaboration with community-based organizations to supplement resources and align efforts. For accountable entities investigating different strategies, activities, and interventions to enhance outcomes related to variations in care, and who are uncertain about the potential impact of those interventions or activities, decision-support tools exist that can help evaluate how new initiatives or policies might affect differences in care among subpopulations.[13]

Furthermore, developers are encouraged to review relevant literature, engage with a TEP, and collaborate with accountable entities to identify proven or potential interventions. This collaborative approach ensures that strategies are not only based on empirical evidence and expert insights but are also effectively tailored to reduce health care differences.

Maintenance Endorsement

For maintenance endorsement, measure developers and stewards should re-examine previously summarized actions to determine if they are still relevant and update accordingly with any new supportive evidence. New actions identified since last endorsement review should also be examined and summarized.

Table 5: Element 4 Common Challenges and Mitigation Approaches

Challenge	Mitigation Approach
Actionable interventions to reduce health care differences have limited evidence or are difficult to identify.	Measure developers should consider the feasibility of improvement efforts as different accountable entities have varying access to resources. Measure developers should develop recommendations based on what is feasible for an accountable entity with minimal resources or take a multi-recommendation approach that considers different levels of resourcing available. The measure developer can also engage a TEP (inclusive of accountable entities) to determine if interventions are actionable. When possible, engage accountable entities in interviews, surveys, or workshops to gather insights on past experiences, challenges faced, and successful strategies. This real-world input is crucial for

---

[13] Olyaeemanesh A, Takian A, Mostafavi H, et al. Health Equity Impact Assessment (HEIA) reporting tool: developing a checklist for policymakers. Int J Equity Health. 2023;22(1):241. Published 2023 Nov 18. doi:10.1186/s12939-023-02031-0

Challenge	Mitigation Approach
	understanding the practical aspects of implementing interventions. Any potential limitations in the actionability of the measure should be noted as part of the element submission. For example, indicate any policy or system changes needed within communities to address identified differences.

Element 4 Example

Initial Endorsement

Encouraged for initial endorsement.

Maintenance Endorsement

To effectively address differences in breast cancer screening recall rates, accountable entities can use the stratification results to identify specific subpopulations who experience higher or lower recall rates and implement tailored strategies to meet their unique needs.

First, entities can enhance radiologist training. Studies have shown that targeted education in recognizing and understanding the nuances of breast tissue density variations among different age and racial groups can help reduce unnecessary recall rates.[14] This specialized training can equip radiologists with the skills needed to make more accurate assessments across diverse populations, thereby reducing the higher recall rates observed.

Improving provider communication, sharing decision-making with patients, increasing the efficiency of follow-up procedures, and tailoring care plans to the cultural and linguistic needs of diverse populations can significantly enhance patient understanding and adherence to follow-up recommendations after abnormal mammograms.[15] Health care providers should ensure that communication materials are clear and accessible and that systems are in place to support consistent follow-up care. This may include implementing reminder systems and providing assistance with scheduling additional screenings, which have been effective in increasing timely follow-up rates among women with dual eligibility.[16]

---

[14] Jones, D. et al. (2021). “Impact of Radiologist Education on Reducing Racial Disparities in Breast Cancer Screening Rates.” Journal of Breast Imaging, 3(2), 123-131.

[15] Williams, R., & Patel, B. (2020). “Cultural Competence in Health Communication: Strategies for Achieving Patient-Centered Care.” Healthcare, 8(1), 15.

[16] Anderson, H. et al. (2021). “Effectiveness of Reminder Systems in Improving Follow-Up Compliance in Breast Cancer Screening Programs.” Journal of Women's Health, 30(5), 678-686.

Lastly, addressing potential biases in AI tools used for mammogram interpretations is essential. As AI technology plays an increasingly prominent role in breast cancer screening, it is important to ensure these tools are trained on diverse datasets to prevent perpetuating existing differences.[17] Providers should collaborate with developers to monitor and adjust AI algorithms, ensuring they do not exacerbate differences in performance across all patient demographics.

By focusing on these areas, health care providers can make significant progress in reducing variations in breast cancer recall rates, leading to improved health outcomes for all women.

For further maintenance reviews, the Element 4 information provided during prior endorsement may serve as a starting point. This section includes additional information that should be provided.

As mentioned in Element 3, no statistically significant changes in the breast cancer recall rates were observed across subpopulations from 2018 to 2023. This suggests that accountable entities may need to reevaluate what kind of improvement work is effective in addressing recall rates. Accountable entities should continue enhancing radiologist training, improving patient communication and follow-up procedures, and addressing potential biases in AI tools used for mammogram interpretations, as previously mentioned study-based actions that can address differences. Accountable entities may also consider engaging local community-based organizations and convening patient advisory panels to inform new approaches for addressing the identified differences.

Conclusion

Evaluating subpopulation differences in the context of health care quality measurement is a crucially important and continually evolving process. This document provides initial guidance to measure developers and stewards for completing the Closing Care Gaps domain, which remains voluntary through the Fall 2026 E&M cycle. Guidance will be updated on a timely basis and in alignment with national priorities to maintain a current reference to assist measure developers and stewards in navigating the Closing Care Gaps domain of the E&M process and in reducing health differences between subpopulations.

---

[17] Kim, J., & Lee, S. (2022). “Evaluating Bias in AI Algorithms for Breast Cancer Detection.” Medical Informatics, 39(1), 45-52.

Appendix A: Resources for Developers

Resource	Description
CMS Measures Management System (MMS) Hub	This online resource is designed to support the development, implementation, and maintenance of quality measures used in CMS health care programs. The MMS Hub serves as a central resource for measure developers, stakeholders, and the public to access information, tools, and guidance related to quality measure development and management.
Endorsement & Maintenance (E&M) Guidebook	This resource is the current guidelines provided by Battelle that support the E&M process. The guidebook focuses on the E&M review process and how to submit measures to Battelle.
Inventory of Resources for Standardized Demographic and Language Data Collection	This document links to resources that can assist in standardizing collection for demographic and language data. It also outlines minimum standards for data collection, best practices and guidelines for standardized data collection, training tools and webinars, and resources that discuss challenges in depth.

Appendix B: Example 2

The maintenance measure [CBE #0133 In-Hospital Risk Standardized Mortality for Percutaneous Coronary Intervention](#) (PCI) is used as an additional example. The example was developed for guidance purposes, does not reflect the actual work of its measure developer/steward, and had no impact on the measure's endorsement decision.

Please note that Elements 2-4 are not fully addressed in this example. The example primarily provides distribution results by defined subpopulations, which is a component of these elements but does not encompass their full scope. Statistical significance tests are needed to assess the differences in measure performance across the defined subpopulations. Readers are encouraged to refer to the main text of the guidance document for comprehensive details on Elements 2-4.

Measure Description: *This measure estimates a hospital-level risk standardized mortality rate (RSMR) in adult patients without cardiogenic shock or cardiac arrest undergoing PCI. The outcome is defined as in-hospital mortality following a PCI procedure performed during the episode of care. Mortality is defined as death from any cause during the episode of care.*

Element 1

Research on subpopulation differences for cardiovascular disease management and outcomes shows racial variations exist in access to PCI, door-to-balloon (DTB) time, procedure utilization, and outcomes among elderly patients. Black and Hispanic patients experienced longer door-to-balloon time and lower rates of PCI utilization compared to white patients.[18] With these comorbidities, "…black race was an independent predictor for worse outcomes, whereas Hispanic ethnicity and Asian race were not."[19] In terms of mortality, research indicates that Black patients had a higher risk of 90-day readmission and cumulative mortality following a PCI compared with white patients. Interestingly, these associations were mediated by "dual eligibility, community economic well-being, and traditional cardiovascular risk factors (e.g., age, sex)."[20] These factors seem to consistently impact Black women more, as Black women still have a higher mortality rate than other racial and ethnic groups from heart disease, stroke, and hypertensive disease. Additionally, the prevalence of modifiable risk factors (such as tobacco

---

[18] Wang, C., Lindquist, K., Krumholz, H., & Hsia, R. Y. (2023). Trends in the likelihood of receiving percutaneous coronary intervention in a low-volume hospital and disparities by sociodemographic communities. PloS one, 18(1), e0279905. https://doi.org/10.1371/journal.pone.0279905

[19] Golomb, M., Redfors, B., Crowley, A., Smits, P. C., Serruys, P. W., Von Birgelen, C., Madhavan, M. V., Ben-Yehuda, O., Mehran, R., Leon, M. B., & Stone, G. W. (2020). Prognostic impact of race in patients undergoing PCI. КАРДИОЛОГИЯ УЗБЕКИСТАНА, 13(13), 1586–1595. https://doi.org/10.1016/j.jcin.2020.04.020

[20] Spehar, S. M., Seth, M., Henke, P., Alaswad, K., Schreiber, T., Berman, A., Syrjamaki, J., Ali, O. E., Bader, Y., Nerenz, D., Gurm, H., & Sukul, D. (2022). Race and outcomes after percutaneous coronary intervention: Insights from the Blue Cross Blue Shield of Michigan Cardiovascular Consortium. American Heart Journal, 255, 106–116. https://doi.org/10.1016/j.ahj.2022.10.001

use, decreased physical activity, and lack of a healthy diet) for cardiovascular disease is higher in Black people than in the general U.S. population.[21]

The literature review reveals that the risk of mortality from PCI is complex and influenced by multiple factors. Although Black individuals represent the racial group most affected by mortality, factors such as dual eligibility, the economic health of the community, and conventional cardiovascular risk factors also play significant roles.

Element 2

In our approach to testing differences in performance scores across identified subpopulations empirically, we:

- Examined variation in hospital performance for the measure based on overall performance and stratified by subpopulations of age, sex, race, ethnicity, and proportion of patients insured through Medicaid to identify if there were any meaningful differences in social risk.
- Ensured that the data used for analysis came only from facilities that met the minimum case count requirements for public reporting. This step was crucial to maintain the reliability and validity of the analysis.
- Examined the distribution of performance scores based on age, sex, race, and insurance.

Element 3

In terms of the overall performance distribution, the median risk-standardized rate of mortality was 0.83%, with an interquartile range of 0.73% to 0.98% (Table B1). Across stratified analysis based on age, sex, patients insured through Medicaid, and proportion of non-white patients, we found significant overlap in the distribution of hospital performance, as detailed below.

Table B1: Distribution of Risk-Standardized Mortality Rates Stratified by Hospital Volume

Description	Hospital Volume	Risk-Standardized Mortality Rates
n	1669	1669
Mean	391.93	0.0088
Std. Deviation	337.88	0.0024
100% Max	3337	0.0282
99%	1541	0.0163
95%	1000	0.0134
90%	803	0.0118

---

[21] Disparities in cardiovascular care: Past, present, and solutions Quentin R. Youmans, Lindsey Hastings-Spaine, Oluseyi Princewill, Titilayo Shobayo, Ike S. Okwuosa Cleveland Clinic Journal of Medicine Sep 2019, 86 (9) 621-632; DOI: 10.3949/ccjm.86a.18088

Description	Hospital Volume	Risk-Standardized Mortality Rates
75% Q3	541	0.0098
50% Median	301	0.0083
25% Q1	157	0.0073
10%	72	0.0064
5%	44	0.0059
1%	14	0.0049
0% Min	2	0.0036

Although most data are centered around the median, there is a slight right skew to the mortality rates (Figure B1).

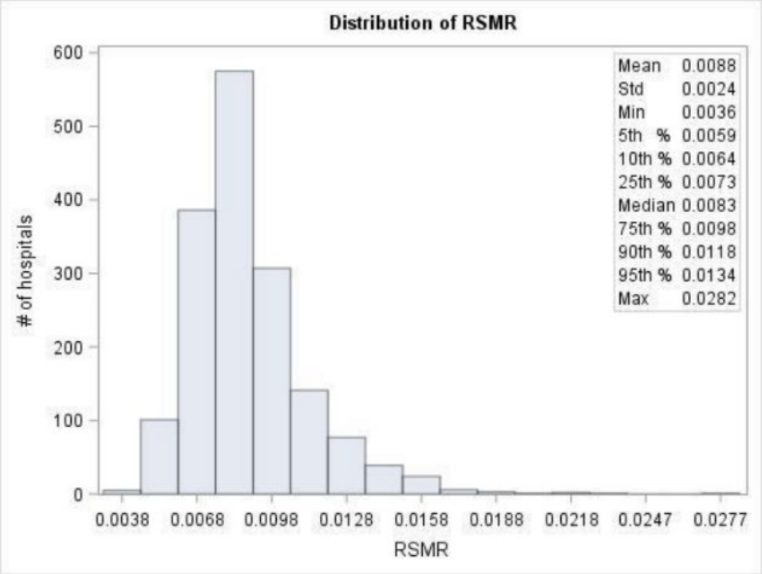


Figure B1: Distribution of Risk-Standardized Mortality Rates

Age

Hospitals (n=1,669) were stratified into quartiles by their proportion of patients over the age of 65 (median: 58.76%, interquartile range [IQR]: 53.71% to 63.93%). Hospital performance was similar across hospitals stratified by quartile based on age (Table B2; Figure B2).

Table B2: Distribution of RSMR Stratified by Quartile of Hospital Percentage Age ≥ 65

Description	%Age≥65	%Age≥65			
		Q1	Q2	Q3	Q4
n	1669	417	417	418	417
Mean	58.35%	0.89%	0.90%	0.87%	0.86%
Std. Deviation	8.84%	0.21%	0.24%	0.26%	0.23%
100% Max	100.00%	1.95%	2.30%	2.82%	1.73%
99%	77.58%	1.61%	1.59%	1.63%	1.64%
95%	71.55%	1.35%	1.37%	1.34%	1.29%
90%	68.35%	1.315%	1.24%	1.19%	1.16%
75% Q3	63.93%	0.97%	1.01%	0.99%	0.96%
50% Median	58.76%	0.84%	0.85%	0.82%	0.81%
25% Q1	53.71%	0.75%	0.74%	0.71%	0.71%
10%	47.80%	0.68%	0.63%	0.60%	0.62%
5%	43.53%	0.66%	0.58%	0.55%	0.55%
1%	32.35%	0.59%	0.49%	0.49%	0.47%
0% Min	0.00%	0.48%	0.42%	0.42%	0.36%

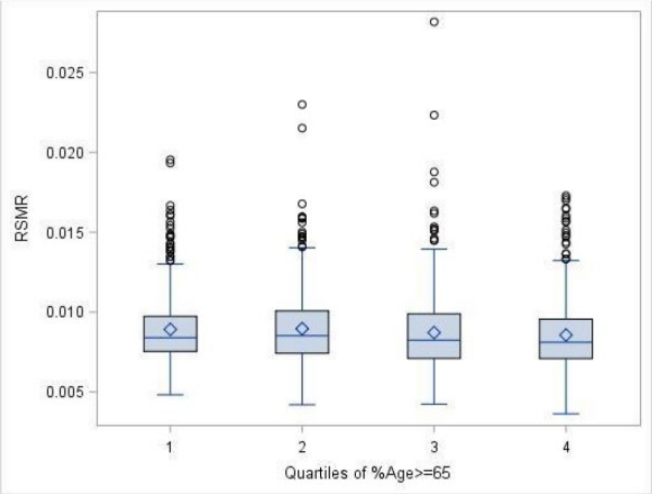


Figure B2: Graphic Distribution of RSMR Stratified by Quartile of Hospital Percentage Age ≥ 65

Sex

The median percent of female patients was 30.74% (IQR: 27.66% to 33.90%), and hospitals performed similarly in RSMR across quartiles of proportion of female patients (Table B3, Figure B3).

Table B3: Distribution of RSMR Stratified by Quartile of Hospital Percent Female

Description	%Female	%Female			
		Q1	Q2	Q3	Q4
n	1669	417	417	417	418
Mean	31.03%	0.86%	0.86%	0.91%	0.89%
Std. Deviation	5.61%	0.22%	0.25%	0.25%	0.22%
100% Max	100.00%	2.82%	2.30%	2.15%	1.93%
99%	46.15%	1.48%	1.62%	1.68%	1.61%
95%	39.32%	1.21%	1.33%	1.41%	1.37%
90%	37.25%	1.12%	1.18%	1.25%	1.17%
75% Q3	33.90%	0.94%	0.98%	1.02%	0.96%
50% Median	30.74%	0.82%	0.81%	0.85%	0.84%
25% Q1	27.66%	0.72%	0.70%	0.73%	0.75%
10%	24.81%	0.64%	0.62%	0.65%	0.66%
5%	23.08%	0.58%	0.56%	0.59%	0.60%
1%	19.95%	0.49%	0.48%	0.52%	0.51%
0% Min	0.00%	0.46%	0.36%	0.48%	0.42%

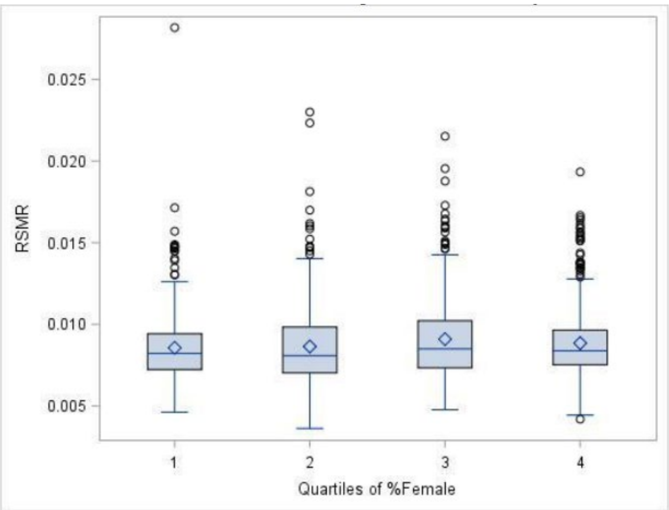


Figure B3: Graphic Distribution of RSMR Stratified by Quartile of Hospital Percent Female

Proportion of Non-White Patients

The median proportion of non-white patients was 12.16% (IQR: 5.46% to 23.18%). Hospital performance across quartiles was similar regardless of the proportion of non-white patients treated, with median performance scores ranging from 0.82% (Q1, lowest proportion of non-white patients) to 0.85% (Q4, highest proportion on non-white patients) (Table B4, Figure B4).

Table B4: Distribution of RSMR Stratified by Quartile of Hospital Percent Non-White

Description	%Non-White	%Non-White			
		Q1	Q2	Q3	Q4
n	1669	417	417	418	417
Mean	16.92%	0.86%	0.86%	0.89%	0.90%
Std. Deviation	15.77%	0.21%	0.21%	0.24%	0.26%
100% Max	96.97%	1.88%	2.82%	1.81%	2.30%
99%	75.00%	1.51%	1.56%	1.64%	1.71%
95%	49.70%	1.32%	1.30%	1.37%	1.41%
90%	37.72%	1.15%	1.14%	1.22%	1.21%
75% Q3	23.18%	0.94%	0.96%	1.01%	1.00%
50% Median	12.16%	0.82%	0.82%	0.85%	0.85%
25% Q1	5.46%	0.73%	0.71%	0.74%	0.73%

Description	%Non-White	%Non-White			
		Q1	Q2	Q3	Q4
10%	2.49%	0.64%	0.63%	0.65%	0.64%
5%	1.49%	0.60%	0.58%	0.59%	0.58%
1%	0.00%	0.52%	0.48%	0.48%	0.53%
0% Min	0.00%	0.44%	0.40%	0.36	0.47%

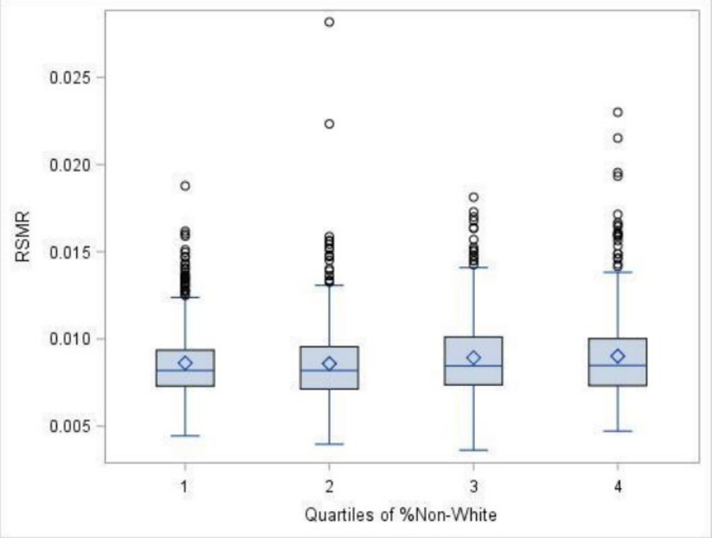


Figure B4: Graphic Distribution of RSMR Stratified by Quartile of Hospital Percent Non-White

Insurance

Hospitals (n=1,669) were stratified into quartiles by their proportion of patients with Medicaid as the primary insurance (median: 11.60%, IQR: 6.80% to 17.41%). Hospital performance was similar across hospitals stratified by quartile based the proportion of patients with Medicaid insurance coverage (Table B5, Figure B5).

Table B5: Distribution of RSMR Stratified by Quartile of Hospital Percent Medicaid

Description	% Medicaid	% Medicaid			
		Q1	Q2	Q3	Q4
n	1669	417	417	418	417
Mean	13.44%	0.87%	0.87%	0.88%	0.89%
Std. Deviation	9.74%	0.21%	0.23%	0.28%	0.22%
100% Max	72.09%	1.93%	1.70%	2.82%	1.95%

Description	% Medicaid	% Medicaid			
		Q1	Q2	Q3	Q4
99%	47.98%	1.59%	1.60%	1.88%	1.59%
95%	31.53%	1.29%	1.33%	1.40%	1.34%
90%	25.42%	1.14%	1.21%	1.20%	1.18%
75% Q3	17.41%	0.97%	0.98%	0.98%	0.99%
50% Median	11.60%	0.83%	0.83%	0.83%	0.84%
25% Q1	6.80%	0.73%	0.73%	0.72%	0.74%
10%	3.53%	0.65%	0.64%	0.61%	0.65%
5%	1.95%	0.60%	0.59%	0.54%	0.59%
1%	0.00%	0.53%	0.49%	0.49%	0.53%
0% Min	0.00%	0.36%	0.42%	0.44%	0.40%

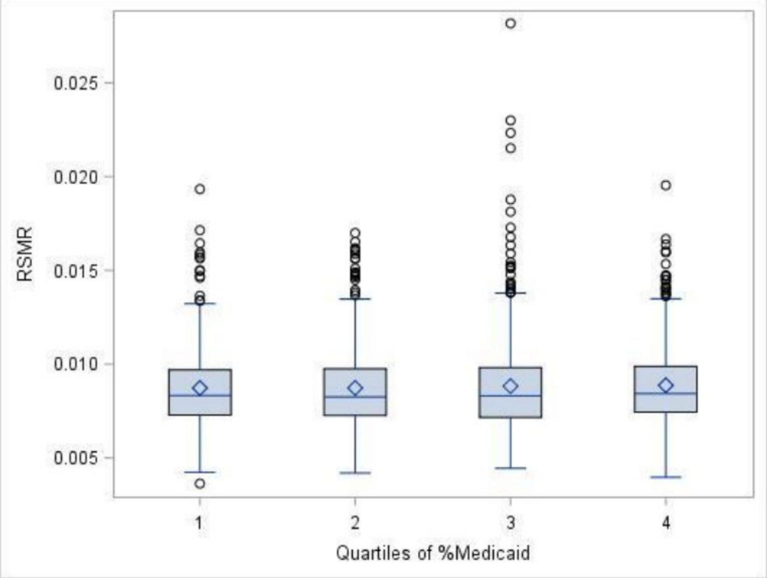


Figure B5: Graphic Distribution of RSMR Stratified by Quartile of Hospital Percent Medicaid

Element 4

Accountable entities are encouraged to benchmark their PCI mortality rates by race, dual eligibility, and any other indicators of SES whenever possible. Some research has shown that specialist intervention and adequate nursing staff helps improve PCI mortality.[22,23] Additionally, entities should proactively seek opportunities to engage with patient communities. This engagement will enhance their understanding of the unique challenges faced by different subpopulations, which will help in effectively addressing performance variations in PCI mortality rates.

---

[22] Masters J, Morton G, Anton I, et al. Specialist intervention is associated with improved patient outcomes in patients with decompensated heart failure: evaluation of the impact of a multidisciplinary inpatient heart failure team. Open Heart. 2017 Mar 8;4(1):e000547. doi: 10.1136/openhrt-2016-000547. PMID: 28409010; PMCID: PMC5384462.

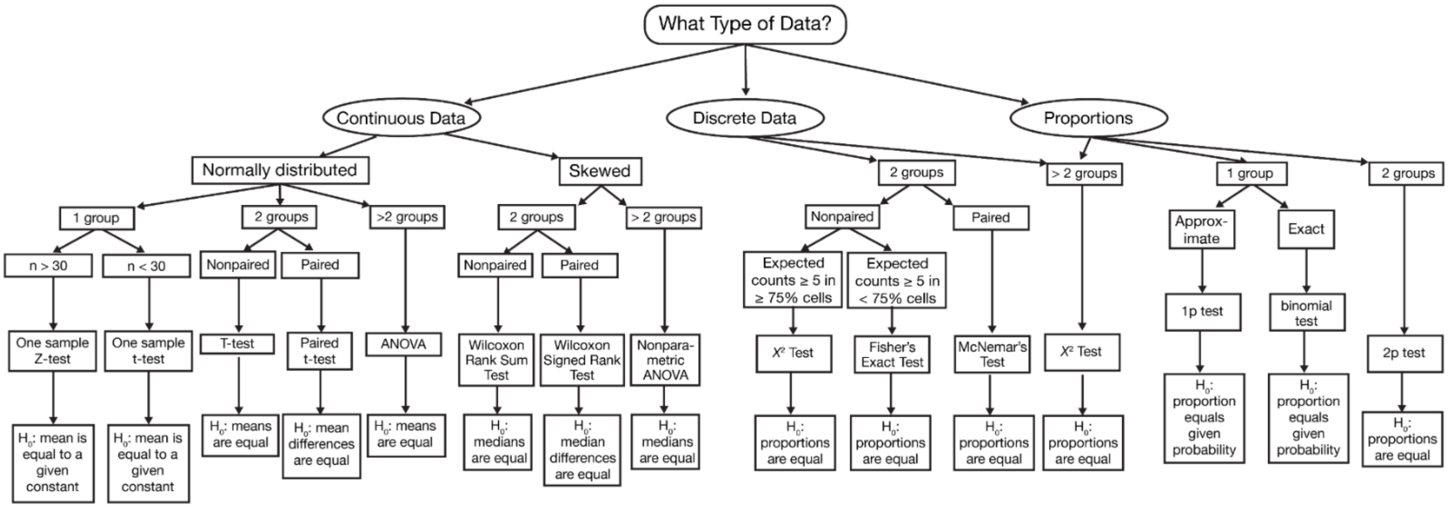
[23] Kim Y, Kim J. In-Hospital Mortality in Patients Receiving Percutaneous Coronary Intervention According to Nurse Staffing Level: An Analysis of National Administrative Health Data. Int J Environ Res Public Health. 2020;17(11):3799. Published 2020 May 27. doi:10.3390/ijerph17113799

Appendix C: Test Statistic Flow Chart

This flow chart from the Osborne Nishimura Lab at Colorado State University may serve as a decision-making tool for developers when considering the suitable statistical test for comparing performance scores across strata. It is not intended to be prescriptive in any way, and developers need to carefully consider their specific data characteristics, look at how the data will be stratified, and choose the most appropriate statistical test. This visual tool simplifies the process of selecting the correct statistical test, ensuring that the analysis is both appropriate and robust, leading to valid and reliable results.

Starting with the type of data at hand, the chart branches into different pathways based on key characteristics such as data distribution (normal or non-normal, i.e., skewed), the number of groups involved, and whether the data are paired or independent. For comparing means between two groups, the chart guides users to choose between a t-test or a non-parametric test, depending on the data's distribution. For comparisons among more than two groups, it directs users to consider analysis of variance (ANOVA) or its non-parametric ANOVA counterpart. Additionally, for categorical data analysis, the chart recommends the chi-square test or Fisher's exact test, based on the sample size and expected frequencies.

Flow chart: which test statistic should you use?



Appendix D: Data Resources

Resource	Description	Type of Data Included
Administrative Claims Data	These are data collected from health care insurance claims, primarily used for billing and reimbursement purposes.	Detailed records of health care services provided to patients, including diagnoses, procedures, patient demographics, provider information, and the cost of services. To directly assess socio-contextual factors, it is often necessary to link claims data with other databases that specifically such factors, such as those from the American Community Survey or Social Vulnerability Index.
American Community Survey (ACS)	Conducted by the U.S. Census Bureau, this survey provides detailed demographic, social, economic, and housing statistics.	Annual data on age, race, income, commute time to work, homeownership, and other important community aspects.
Area Deprivation Index (ADI)	This index ranks neighborhoods by socioeconomic status disadvantage.	Metrics on income, education, employment, and housing quality to assess neighborhood-level deprivation.
Behavioral Risk Factor Surveillance System (BRFSS)	This system collects state data about U.S. residents' health-related risk behaviors.	Information on health risk behaviors, chronic health conditions, and use of preventive services.
Electronic Health Records (EHRs)	EHRs increasingly include socio-contextual data, which can be used for stratification.	Information on social factors such as housing stability, food security, transportation, and social connections.
Healthcare Cost and Utilization Project (HCUP)	HCUP databases provide data on hospital care, including patient demographics.	Comprehensive hospital care data including patient demographics, diagnoses, procedures, and charges.
Social Vulnerability Index (SVI)	Developed by the Centers for Disease Control and Prevention (CDC), this index helps to identify socially vulnerable populations.	Fifteen U.S. census variables that help identify the social vulnerability of every census tract.

Appendix E: Acronyms

Please note: The following list encompasses acronyms that Battelle commonly encounters and uses in its work as a CBE. Not all the acronyms will appear in this document.

Acronym	Definition
ACA	Affordable Care Act
ACC	American College of Cardiology
ACO	Accountable Care Organization
AGC	After Government Contract
AHIP	Formerly known as American Health Insurance Partnership
AHRQ	Agency for Healthcare Research and Quality
AI Pilot	Artificial Intelligence Pilot
AIPAC	Advanced Illness and Post-Acute Care
AIR	American Institutes for Research
ANOVA	Analysis of Variance
ASCO	American Society of Clinical Oncology
ASCQR	Ambulatory Surgical Center Quality Reporting Program
ASCs	Ambulatory Surgical Centers
C&E	Cost and Efficiency
CAH	Critical Access Hospital
CAHPS	Consumer Assessment of Healthcare Providers and Systems
CBE	Consensus-Based Entity
CBE ID	Consensus-Based Entity Identification
CDC	Centers for Disease Control and Prevention
CDS	Clinical Decision Support
CDSS	Clinical Decision Support System
CIS	Clinical Information Systems
CMIT	CMS Measures Inventory Tool
CMMI	Center for Medicare and Medicaid Innovation
CMS	Centers for Medicare & Medicaid Services
CO	Contracting Officer
COIs	Conflicts of Interest
COR	Contracting Officer's Representative
CPG	Clinical Practice Guidelines

Acronym	Definition
CQL	Clinical Quality Language
CQM	Clinical Quality Measure
CQMC	Core Quality Measures Collaborative
CSAC	Consensus Standards Approval Committee
DEL	CMS Data Element Library
Del.	Deliverable
DOI	Disclosure of Interest
dQMs	Digital Quality Measures
DRC	Direct Reference Code
E&M	Endorsement and Maintenance
EC	Electronic Copy
eCQI	Electronic Clinical Quality Improvement
eCQM	Electronic Clinical Quality Measures
ED	Emergency Department
EHR	Electronic Health Record
EPC	Evidence-Based Practice Center
ESRD QIP	End-Stage Renal Disease Quality Improvement Program
EVI	Expected Value of Information
FAQs	Frequently Asked Questions
FFS	Fee-For-Service
FHIR	Fast Healthcare Interoperability Resources
FMS	Full Measure Submission
FY	Fiscal Year
HACRP	Hospital-Acquired Conditions Reduction Program
HCBS	Home and Community-Based Services
HCD	Human-Centered Design
HEDIS	Healthcare Effectiveness Data and Information Set
HH QRP	Home Health Quality Reporting Program
HH VBP	Home Health Value-Based Purchasing
HHS	Department of Health and Human Services
HIQR	Hospital Inpatient Quality Reporting
HOPD	Hospital Outpatient Department
HOPE	Hospice Outcomes and Patient Evaluation

Acronym	Definition
HOQR	Hospital Outpatient Quality Reporting
HQMF	Health Quality Measurement Format
HQR	Hospice Quality Reporting
HQRP	Hospice Quality Reporting Program
HRRP	Hospital Readmission Reduction Program
HSAG	Health Services Advisory Group
HTML	Hypertext Markup Language
HVBP	Hospital Value-Based Purchasing
IAW	In Accordance With
ICD	International Classification of Diseases (International Statistical Classification of Diseases and Related Health Problems)
IHI	Institute for Healthcare Improvement
IMPACT Act	Improving Medicare Post-Acute Care Transformation Act
IPF	Inpatient Psychiatric Facilities
IPF QRP	Inpatient Psychiatric Facility Quality Reporting Program
IPPS	Inpatient Prospective Payment System
IQR	Inpatient Quality Reporting
IR	Initial Recognition
IRF	Inpatient Rehabilitation Facilities
IRF QRP	Inpatient Rehabilitation Facility Quality Reporting Program
IT	Information Technology
ITS	Intent to Submit
LLMs	Large Language Models
LTACH	Long-Term Acute Care Hospitals
LTCH	Long-Term Care Hospital
LTCH QRP	Long-Term Care Hospital Quality Reporting Program
MA	Medicare Advantage
MACRA	Medicare Access and CHIP Reauthorization Act
MACS	Medicaid: Adult Core Set
MAQIP	Medicare Advantage Quality Improvement Program
MAT	Measure Authoring Tool
MCCS	Medicaid: Child Core Set
MCO	Managed Care Organization

Acronym	Definition
MERIT	Measures Under Consideration Entry/Review Tool
MIPPA	Medicare Improvement for Patients and Providers Act of 2008
MIPS	Merit-based Incentive Payment System
MLTSS	Managed Long-Term Service and Support
MMS	Measures Management System
MS-DOI	Measure-Specific Disclosure of Interest
MSR	Measure Set Review
MSSP	Medicare Shared Savings Program
MUC	Measures Under Consideration
n	Sample Size
NCDC	National Consensus Development and Strategic Planning for Health Care Quality Measurement Contract
NCQA	National Committee for Quality Assurance
NHDNG	Novel Hybrid Delphi and Nominal Groups
NHQI	Nursing Home Quality Initiative
NLP	Natural Language Processing
NQF	National Quality Forum
NQS	CMS National Quality Strategy
NTTAA	National Technology Transfer and Advancement Act
OMB	Office of Management and Budget
OP	Option Period
OY	Option Year
PA	Preliminary Assessment
PAC/LTC	Post-Acute Care/Long-Term Care
PaLS	Patient Life Goals Survey
PAM	Patient Activation Measure
PCHQR	PPS-Exempt Cancer Hospital Quality Reporting
PDF	Portable Document Format
PIE Form	Pre-Meeting Initial Evaluation Form
PL	Project Leader
PM	Project Manager
PMP	Project Management Plan
POC	Point of Contact

Acronym	Definition
PPS	Prospective Payment System
PQA	Pharmacy Quality Alliance
PQM	Partnership for Quality Measurement
PRA	Paperwork Reduction Act
PRMR	Pre-Rulemaking Measure Review
PRO	Patient-Reported Outcome
PROM	Patient-Reported Outcome Measure
PRO-PMs	Patient-Reported Outcome Performance Measures
Q&A	Question & Answer
QC	Quality Control
QCDR	Qualified Clinical Data Registries
QDM	Quality Data Model
QI	Quality Improvement
QMDSA	Quality Measure Developer and Steward Agreement
QPP	Quality Payment Program
REHQR	Rural Emergency Hospital Quality Reporting (Program)
SES	Socioeconomic Status
SLIN	Subline Item Number
SMEs	Subject Matter Experts
SMP	Scientific Measures Panel
SNF	Skilled Nursing Facilities
SNF QRP	Skilled Nursing Facility Quality Reporting Program
SNF VBP	Skilled Nursing Facility Value-Based Purchasing
SOP	Standard Operating Procedure
SOW	Statement of Work
SSA	Social Security Administration
STAR	Submission Tool and Repository
SUD	Substance Use Disorder
TBD	To Be Determined
TEP	Technical Expert Panel
TL	Task Lead
UMLS	Unified Medical Language System
USCDI	United States Core Data for Interoperability

Acronym	Definition
VSAC	Value Set Authority Center
Yale CORE	Yale Center for Outcomes Research and Evaluation

Acknowledgements

Partnership for Quality Measurement Organizations

Battelle

Institute for Healthcare Improvement (IHI)

Guidance Authors

Battelle

Matthew K. Pickering (PharmD), Principal Quality Measure Scientist

Anna Michie (MHS, PMP), Senior Quality Measure Scientist

IHI

This guidance was developed in partnership with the IHI. Primary contributors include:

Maddie Little-Ghose, MPH, Project Manager

Kaely Burgess, BS, Associate Project Manager

Special thanks to the additional IHI staff who consulted on drafting and strengthening this guidance: Marianne Smith, MAS-PSHQ, BSN, RN, CPPS; and Jesse McCall, MBA.

Special thanks to the IHI Equity Faculty who consulted on drafting and strengthening this guidance: Kim Herner, MD; Kendall LaSane, DrPH, MPH; Kristen Azar, PhD, RN, MSN/MPH FAHA; and Ron Wyatt, MD, MHA.

PQM Partnership for
Quality Measurement
Powered by Battelle

Version 1.0 | July 2025 | *The analyses upon which this publication (or document) 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. Restricted:* Use, duplication, or disclosure is subject to the restrictions as stated in Contract Number 75FCMC23C0010 between the Government and Battelle.