

Additional Testing for Breast Cancer Screening Episode-Based Cost Measure

This document includes testing results at the TIN-NPI level and a summary of empirical data that demonstrates support for the measure concept.

TIN-NPI Level Testing Results

Reliability and Validity

Subsection	Row	Field Label	Guidance	ADD YOUR CONTENT HERE
Measure Score Level (Accountable Entity Level) Testing	032	*Reliability	<p>Indicate whether reliability testing was conducted for the accountable entity-level measure scores. Acceptable reliability tests include signal-to-noise (or inter-unit reliability) or random split-half correlation. For more information on accountable entity-level reliability testing, refer to the Blueprint content on the CMS Measures Management System (MMS) Hub (https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria/scientific-acceptability/reliability).</p> <p>Select “Yes” if acceptable accountable entity-level reliability testing has been completed as of submission of this form.</p> <p>Select “No” if you are not able to provide the results of acceptable accountable entity-level reliability testing in this submission. If testing results are incomplete, or if you are submitting a different type of reliability testing, provide as an attachment.</p> <p>Note: This section refers to the reliability of the accountable entity-level measure scores in the final performance measure. For testing of surveys or patient reported tools, refer to the Patient-Reported Data section. Note: for MIPS-Quality submissions, please provide individual clinician-level results. If the measure was also tested at the clinician group level, you may include those results in an attachment.</p>	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

Measure Score Level (Accountable Entity Level) Testing	033	*Reliability: Type of analysis	<p>Select all that apply.</p> <p>Signal-to-noise (or inter-unit reliability) is the precision attributed to an actual construct versus random variation (e.g., ratio of between unit variance to total variance) (Adams J. The reliability of provider profiling: a tutorial. Santa Monica, CA: RAND; 2009. http://www.rand.org/pubs/technical_reports/TR653.html).</p> <p>Random split-half correlation is the agreement between two measures of the same concept, using data derived from split samples drawn from the same entity at a single point in time.</p>	<input checked="" type="checkbox"/> Signal-to-Noise <input type="checkbox"/> Random Split-Half Correlation
Measure Score Level (Accountable Entity Level) Testing	034	*Signal-to-Noise: Level of Analysis	<p>Select the level of analysis at which the signal-to-noise analysis was conducted. If the measure is specified and intended for use at more than one level, ensure the results in this section are at the same level of analysis selected in the Measure Information section of this form.</p> <p>For MIPS-Quality submissions, you must report the results of individual clinician-level testing. If group-level testing is available, you may submit those results as an attachment.</p>	<input type="checkbox"/> Accountable Care Organization <input type="checkbox"/> Clinician – Group <input checked="" type="checkbox"/> Clinician – Individual <input type="checkbox"/> Facility <input type="checkbox"/> Health plan <input type="checkbox"/> Integrated Delivery System <input type="checkbox"/> Medicaid program (e.g., Health Home or 1115) <input type="checkbox"/> Population: Community, County or City <input type="checkbox"/> Population: Regional and State
Measure Score Level (Accountable Entity Level) Testing	035	*Signal-to-Noise: Sample size	<p>Indicate the number of accountable entities sampled to test the final performance measure. Note that this field is intended to capture the number of measured entities and not the number of individual patients or cases included in the sample.</p>	16,289
Measure Score Level (Accountable Entity Level) Testing	036	*Signal-to-Noise: Median Statistical result	<p>Indicate the median result for the signal-to-noise analysis used to assess accountable entity level reliability. Results should range from 0.00 to 1.00. Calculate reliability as the measure is intended to be implemented (e.g., after applying minimum denominator requirements, appropriate type of setting, provider, etc.).</p>	0.929

Measure Score Level (Accountable Entity Level) Testing	037	*Signal-to-Noise: Interpretation of results	Describe the type of statistic and interpretation of the results (e.g., low, moderate, high). Provide the distribution of signal-to-noise results across measured entities (e.g., min, max, percentiles). List accepted thresholds referenced and provide a citation. If applicable, include the precision of the statistical result (e.g., 95% confidence interval) and/or an assessment of statistical significance (e.g., p-value).	At the testing volume of 20 episodes, the reliability score for the Breast Cancer Screening episode-based cost measure is high, specifically 0.929 at the TIN-NPI level. CMS generally considers 0.4 as the threshold indicating 'moderate' reliability and 0.7 indicating 'high' reliability, which is supported by previous work into reliability and the threshold was finalized in the 2022 Physician Fee Schedule final rule.
Measure Score Level (Accountability Entity Level) Testing	042	*Empiric Validity	<p>Indicate whether empiric validity testing was conducted for the accountable entity-level measure scores. For more information on accountable entity level empiric validity testing, refer to the Blueprint content on the CMS MMS Hub (https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria/scientific-acceptability/validity)</p> <p>Note: This section refers to the empiric validity of the accountable entity level measure scores in the final performance measure. Refer to the Patient-Reported Data section for testing of surveys or patient reported tools.</p> <p>Note: for MIPS-Quality submissions, please provide individual clinician-level results. If the measure was also tested at the clinician group level, you may include those results in an attachment.</p>	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No
Measure Score Level (Accountable Entity Level) Testing	043	*Empiric Validity: Level of Analysis	<p>Select the level of analysis at which the empiric validity analysis was conducted. If the measure is specified and intended for use at more than one level, ensure the results in this section are at the same level of analysis selected in the Measure Information section of this form.</p> <p>For MIPS-Quality submissions, you must report the results of individual clinician-level testing. If group-level testing is available, you may submit those results as an attachment.</p>	<input type="checkbox"/> Accountable Care Organization <input type="checkbox"/> Clinician – Group <input checked="" type="checkbox"/> Clinician – Individual <input type="checkbox"/> Facility <input type="checkbox"/> Health plan <input type="checkbox"/> Integrated Delivery System <input type="checkbox"/> Medicaid program (e.g., Health Home or 1115) <input type="checkbox"/> Population: Community, County or City <input type="checkbox"/> Population: Regional and State
Measure Score Level (Accountability Entity Level) Testing	044	*Empiric Validity: Sample size	Indicate the number of accountable entities sampled to test the final performance measure. Note that this field is intended to capture the number of measured entities and not the number of individual patients or cases included in the sample.	16,289

<p>Measure Score Level (Accountability Entity Level) Testing</p>	<p>045</p>	<p>* Empiric Validity: Methods and findings</p>	<p>Describe the methods used to assess accountable entity level validity. Describe the comparison groups or constructs used to verify the validity of the measure scores, including hypothesized relationships (e.g., expected to be positively or negatively correlated). Describe your findings for each analysis conducted, including the statistical results and the strongest and weakest results across analyses. If applicable, include the precision of the statistical result(s) (e.g., 95% confidence interval) and/or an assessment of statistical significance (e.g., p-value). If methods and results require more space, include as an attachment.</p>	<p>This measure is tested using a mediation analysis to demonstrate construct validity and a correlation analysis to demonstrate concurrent validity.</p> <p>The mediation analysis estimates both the direct and indirect effect of treatment choices on the measure score. This analysis first estimates the correlation between treatment choices and the measure score while controlling for adverse outcomes. Then the correlation between treatment choices and related adverse outcomes is calculated to demonstrate the indirect effect. Generally, adverse outcomes are non-trigger inpatient hospitalizations, non-trigger emergency room visits, and post-acute care. The remaining service categories are typically considered treatment. The results show that spending on ambulatory/minor procedures, anesthesia, and laboratory testing is statistically associated with a better measure score. On the other hand, the main drivers of cost are major procedures, outpatient evaluation and management, and imaging services. While major procedures show a statistical association with lower costs of adverse events, the reduction in costs of adverse events is not enough to offset the cost of major procedures, which further reinforces the need for early detection to avoid high-intensity treatments. Even though outpatient evaluation and management and imaging are essential, the results indicate that they can be prone to overuse because more spending on these services does not offset the costs of adverse events.</p> <p>The correlation analysis shows that the cost measure score is positively associated with the Breast Cancer Screening Recall Rate measure ($r = 0.27$, $p\text{-value} < 0.001$) and OP-39 Breast Cancer Screening Recall Rate ($r = 0.21$, $p\text{-value} < 0.001$), which is aligned with mediation analysis in suggesting that imaging services are necessary but can be prone to overuse. The cost measure score is negatively associated with Breast Cancer Screening with an Eventual Breast Cancer Diagnosis: PPV1 measure ($r = -0.13$, $p\text{-value} < 0.001$)</p>
--	------------	---	--	---

				and Use of Biopsy After Diagnostic Follow-up with an Eventual Breast Cancer Diagnosis: PPV3 ($r = -0.14$, p -value <0.001), which are also aligned in with the mediation analysis in emphasizing the importance of cancer detection in reducing costs of delayed treatment or adverse events.
Measure Score Level (Accountable Entity Level) Testing	046	*Empiric Validity: Interpretation of results	Indicate whether the statistical result affirmed the hypothesized relationship for the analysis conducted.	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No
Measure Score Level (Accountable Entity Level) Testing	047	*Face validity	Indicate if a vote was conducted among experts and patients/caregivers on whether the final performance measure scores can be used to differentiate good from poor quality of care. Select "No" if experts and patients/caregivers did not provide feedback on the final performance measure at the specified level of analysis or if the feedback was related to a property of the measure unrelated to its ability to differentiate performance among measured entities. This item is intended to assess whether face validity testing was conducted on the final performance measure and is not intended to assess whether patient-reported surveys or tools have face validity. Survey item testing results can be provided in an attachment and described in the Patient-Reported Data Section.	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No
Measure Score Level (Accountable Entity Level) Testing	048	*Face validity: Total number of voting experts and patients/caregivers	Indicate the number of experts and patients/caregivers who voted on face validity (specifically, whether the measure could differentiate good from poor quality care among accountable entities).	8
Measure Score Level (Accountable Entity Level) Testing	049	*Face validity: Number of experts and patients/caregivers who voted in agreement	Indicate the number of experts and patients/caregivers who voted in agreement that the measure could differentiate good from poor quality care among accountable entities. If votes were conducted using a scale, sum all responses in agreement with the statement. Do not include neutral votes. If more than one question was asked of the experts and patients/caregivers, only provide results from the question relating to the ability of the final performance measure to differentiate good from poor quality care.	7

Measure Score Level (Accountable Entity Level) Testing	050	Face validity: Interpretation	Briefly explain the interpretation of the result, including any disagreement with the face validity of the performance measure.	There were six votes for agree (75%), one vote for strongly agree (12.5%), and one vote for undecided (12.5%) about whether the cost measure could distinguish good from poor quality care.
--	-----	-------------------------------	---	---

Measure Performance Scores and Performance Gap Analysis

Subsection	Row	Field Label	Guidance	ADD YOUR CONTENT HERE
Measure Performance	060	* Measure performance - type of score	Select one. Measure performance score type should be at the level of accountable entity.	<input type="checkbox"/> Categorical (e.g., measured entity scores yes/no, pass/fail, or rating scale/score) <input type="checkbox"/> Composite scale/non-weighted score <input type="checkbox"/> Composite scale/weighted score <input checked="" type="checkbox"/> Continuous variable (e.g., average) <input type="checkbox"/> Count <input type="checkbox"/> Frequency Distribution <input type="checkbox"/> Proportion <input type="checkbox"/> Rate <input type="checkbox"/> Ratio
Measure Performance	061	* Measure performance score interpretation	Select one	<input type="checkbox"/> Better quality = Higher score <input checked="" type="checkbox"/> Better quality = Lower score <input type="checkbox"/> Better quality = Score within a defined interval <input type="checkbox"/> Passing score above a specified threshold defines better quality <input type="checkbox"/> Passing score below a specified threshold defines better quality
Measure Performance	062	* Number of accountable entities included in analysis	Provide the number of accountable entities included in the analysis of the distribution of performance scores. Please enter a single value and do not enter a range. If unknown or not available, enter 9999.	16,289
Measure Performance	063	* Number of accountable entities: unit	Provide the unit of accountable entities included in the analysis of the distribution of performance scores.	Individual clinician (TIN-NPIs) with at least 20 attributed episodes
Measure Performance	064	* Number of persons	Provide the number of persons included in the analysis of the distribution of performance scores	4,694,283

Subsection	Row	Field Label	Guidance	ADD YOUR CONTENT HERE
Measure Performance	065	*10th percentile	<p>Provide the performance score at the 10th percentile for the testing sample that is relevant to the intended use of the measure.</p> <p>If this is a proportion measure, provide the 10th percentile score in percentage form, without the symbol. For example, if the 10th percentile performance score is 21.2%, enter 21.2 and not 0.212.</p> <p>If a 10th percentile performance score is not available, enter 9999.</p>	\$205.98
Measure Performance	066	*50th percentile (median)	<p>Provide the median performance score (50th percentile) for the testing sample that is relevant to the intended use of the measure.</p> <p>Please enter only one value in the response field and do not enter a range of values.</p> <p>If this is a proportion measure, provide the median performance score in percentage form, without the symbol. For example, if the median performance score is 85.6%, enter 85.6 and not 0.856.</p> <p>If a median performance score is not available, enter 9999.</p>	\$250.70
Measure Performance	067	*90th percentile	<p>Provide the performance score at the 90th percentile for the testing sample that is relevant to the intended use of the measure.</p> <p>If this is a proportion measure, provide the 90th percentile score in percentage form, without the symbol. For example, if the 90th percentile performance score is 85.6%, enter 85.6 and not 0.856.</p> <p>If a 90th percentile performance score is not available, enter 9999.</p>	\$294.59

Subsection	Row	Field Label	Guidance	ADD YOUR CONTENT HERE
Measure Performance	068	* Additional measure performance information	<p>Provide the following additional measure performance information, <u>as applicable</u>:</p> <ul style="list-style-type: none"> - Mean performance score across accountable entities in the test sample that is relevant to the intended use of the measure. - Minimum and maximum performance score for the testing sample that is relevant to the intended use of the measure. - Standard deviation of performance scores for the testing sample that is relevant to the intended use of the measure. - Passing score for the performance measure. - Performance score's defined interval, including upper and lower limit of the performance score. 	<ul style="list-style-type: none"> - Mean performance score: \$250.17 - Minimum performance score: \$101.31 - Maximum performance score: \$586.60 - Standard deviation of performance scores: \$40.06
Measure Performance	069	* Is there evidence for statistically significant gaps in measure score performance among select subpopulations of interest defined by one or more social risk factors?	<p>Select one. Social risk factors may include age, race, ethnicity, linguistic and cultural context, sex, gender, sexual orientation, social relationships, residential and community environments, Medicare/Medicaid dual eligibility, insurance status (insured/uninsured), urbanicity/rurality, disability, and health literacy.</p>	<ul style="list-style-type: none"> <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Not tested

Summary of Empirical Data Supporting the Measure Concept

Women have a 1 in 8 chance of developing breast cancer during their life.¹ Breast cancer accounts for around 30% of all new cancers for women each year. It is estimated that in 2022, there will be approximately 287,850 new cases of invasive breast cancer diagnosed and 43,250 deaths from breast cancer.^{1,1} Breast cancer found during screening, before symptoms appear, is less likely to spread, including beyond the breast (metastasis). Early detection makes it easier to treat breast cancer successfully, with a better prognosis for the patient. Screening mammography reduces breast cancer mortality by an estimated 20%-35% in women aged 50-69 years.² As such, early detection is one of the most important strategies for preventing deaths from breast cancer, the second leading cause of cancer death in women in the United States.¹

The literature scan identified three critical areas for improving care and reducing costs. These include improving screening and diagnostic accuracy, incentivizing early cancer detection, and reducing unnecessary resource use. The estimated cost for mammography screening in the US in 2010 was \$7.8 billion.³ While costs associated with appropriate use are not concerning, costs associated with excessive use (e.g., unnecessary repeat imaging) or delayed detection are a significant concern. For example, costs may be high for false-positives and increased follow-up visits.⁴ It is estimated that national expenditure for these false-positive follow-ups cost around \$4 billion a year.⁵ Increasing the accuracy of screenings can lower the need for excessive and expensive follow up treatment. In addition, one study showed that the annual estimated cost for breast cancer screening for women ages 40-49 was \$2.13 billion despite unclear benefits for women in this age group receiving screenings.⁶

Moreover, with approximately 6.1 million screening mammograms in Medicare Part B physician/supplier billing, a modest improvement among these radiologists to recall 20% fewer patients would result in up to 92,000 fewer recalls. This would represent roughly \$12.7 million in savings (average Medicare allowed amount of \$140).⁷ Delays in detection of breast cancer are expected to contribute to higher costs of treating cancer once detected. One review of the literature on breast cancer treatment costs found the stage at initial diagnosis to be an important determinant of resource use. For instance, cancers diagnosed at stage 2 were

¹ American Cancer Society. "American cancer society recommendations for the early detection of breast cancer." ACS Breast Cancer Early Detection Recommendations (2022).

² Elmore JG, Armstrong K, Lehman CD, Fletcher SW. Screening for breast cancer. *JAMA*. 2005;293(10):1245-1256. doi:10.1001/jama.293.10.1245.

³ O'Donoghue, Cristina, Martin Eklund, Elissa M. Ozanne, and Laura J. Esserman. "Aggregate cost of mammography screening in the United States: comparison of current practice and advocated guidelines." *Annals of internal medicine* 160, no. 3 (2014): 145-153.

⁴ Morris, Elizabeth, Stephen A. Feig, Madeline Drexler, and Constance Lehman. "Implications of overdiagnosis: impact on screening mammography practices." *Population health management* 18, no. S1 (2015): S-3.

⁵ Ong, Mei-Sing, and Kenneth D. Mandl. "National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at \$4 billion a year." *Health affairs* 34, no. 4 (2015): 576-583.

⁶ Kunst, Natalia, Jessica B. Long, Xiao Xu, Susan H. Busch, Kelly A. Kyanko, Ilana B. Richman, and Cary P. Gross. "Use and costs of breast cancer screening for women in their 40s in a US population with private insurance." *JAMA internal medicine* 180, no. 5 (2020): 799-801.

⁷ Centers for Medicare & Medicaid Services. Medicare provider utilization and payment data: physician and other practitioners. Accessed May 15, 2022. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier>

determined to have treatment costs that were 32 percent higher than cancers diagnosed at stage 1. Breast cancers detected in stage 3 and stage 4 were, respectively, found to cost 95% and 109% more than cancers detected in stage 1.⁸

Existing literature and Acumen’s preliminary testing indicate a high cost to Medicare for breast cancer screening, opportunities for improvement through best practices, and a substantial empirical performance gap. Our testing indicates that the Breast Cancer Screening measure would have a significant impact on Medicare, whether through measuring beneficiaries, clinicians, or cost. The Breast Cancer Screening cost measure would capture 4,694,293 beneficiaries (using 2022 as the study year). Table 1 shows the distribution of the measure score for TIN and TIN-NPI levels. Substantial variation is observed in the measure, with wide variability between the 90th percentile score and 10th percentile score at the TIN (\$292.11 vs. \$196.56) and TIN-NPI (\$294.59 vs. \$205.98) level. The results suggest that there is an opportunity for improvement in performance across clinicians.

Table 1. Distribution of the Measure Score

Metric	TIN	TIN-NPI
Mean Score	\$247.30	\$250.17
Score Interquartile Range (IQR)	\$50.80	\$49.58
Standard Deviation	\$43.75	\$40.06
Coefficient of Variation	0.18	0.16
Minimum Score	\$120.87	\$101.31
Maximum Score	\$587.42	\$586.60
Score Percentile		
10 th	\$196.56	\$205.98
25 th	\$221.20	\$223.79
50 th	\$247.62	\$250.70
75 th	\$272.00	\$273.38
90 th	\$292.11	\$294.59

⁸ Sun L, Legood R, Santos-Silva I, et al. Global Treatment Costs of Breast Cancer by Stage: A Systematic Review. PLoS One, 2018; 13(11): 30207993.