

Breast Cancer Screening

Measure Testing Form

January 2026



Contents

1.0	Introduction.....	3
1.1	Project Title and Overview.....	3
1.2	Measure Name.....	3
1.3	Type of Measure	3
1.4	Data.....	3
2.0	Preliminary Testing Results	4
2.1	Measure Coverage.....	4
2.2	Frequently Attributed Specialties	5
2.3	Reliability	5
2.4	Validity	6
	2.4.1 Conceptual Model.....	6
	2.4.2 Empirical Validity Analyses.....	8
2.5	Performance Gap.....	9
2.6	Risk Adjustment and Stratification	10
	2.6.1 Discrimination.....	10
	2.6.2 Calibration.....	11
2.7	Dual Status and Provider Location Analysis.....	11
2.8	Impact of Exclusions	14
	Appendix A. Glossary	17
	Appendix B. Distributions of Measure Score	19

1.0 Introduction

This Measure Testing Form (MTF) provides a summary of the preliminary measure testing results as part of field testing the Breast Cancer Screening episode-based cost measure. Field testing, also known as beta testing, is part of the measure development and testing process outlined by the Centers for Medicare & Medicaid Services (CMS) Measures Management System (MMS).¹ This testing occurs after development of initial technical specifications and is an opportunity to assess scientific acceptability (e.g., the extent to which the measure produces valid and reliable results about the intended area of measurement) and usability of a measure (e.g., whether people or organizations can effectively use a measure). Field testing also provides additional evidence to support importance and feasibility evaluations. More information about measure evaluation criteria is available on the MMS Hub.²

Readers may review aggregate field testing results in this document, alongside other measure documentation and field testing reports, to provide feedback on the draft measure using the [field testing survey](#). The testing results reflect the performance of the measure as specified at the time of field testing, which is part of the measure development process. Please see the Draft Cost Measure Methodology for a description of the measure specifications and the Draft Measure Codes List for the list of codes used to specify the measure.³

1.1 Project Title and Overview

CMS has contracted with Acumen, LLC to develop care episode and patient condition groups for use in cost measures to meet the requirements of the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA). The contract name is “Physician Cost Measures and Patient Relationship Codes (PCMP).” The contract number is 75FCMC18D0015, Task Order 75FCMC24F0142.

1.2 Measure Name

Breast Cancer Screening Episode-Based Cost Measure

1.3 Type of Measure

Cost/Resource Use

1.4 Data

The study period is from January 1, 2024, through December 31, 2024. All episodes ending during the study period that meet inclusion and exclusion criteria are included in testing. The measure is calculated with Medicare Parts A and B administrative claims data, the Long-Term Minimum Data Set, and the Medicare Enrollment Database.

Testing results are presented at a testing volume threshold of 10 episodes for clinician groups and individual clinicians. Throughout this document, 'Group' refers to clinician groups identified by a Tax Identification Number (TIN), and 'Clinician' refers to individual clinicians identified by a combination of a TIN and National Provider Identifier (TIN-NPI).

¹ MMS, “Measure Testing Process,” <https://mmshub.cms.gov/measure-lifecycle/measure-testing/process/overview>

² MMS, “Measure Evaluation Criteria,” <https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria>

³ These documents will be available on the CMS Cost Measures Information page once field testing begins: <https://www.cms.gov/medicare/quality/value-based-programs/cost-measures/current>.

2.0 Preliminary Testing Results

This section presents preliminary testing results based on the measure as specified for field testing.

Section 2.1 provides an overview of the measure's coverage of beneficiaries. Section 2.2 lists the most frequently attributed specialties. Sections 2.3 presents reliability testing results. Section 2.4 includes conceptual and empirical validity information. Section 2.5 demonstrates the measure's performance gap. Section 2.6 presents empirical results of the risk adjustment and stratification methods used by this measure. Section 2.7 examines the associations between non-clinical factors—including dual Medicare and Medicaid enrollment status and provider location (urban versus rural)—and measure performance. Section 2.8 examines the impact of exclusion criteria used by the measure through their frequency and resource use patterns. Appendix A provides a glossary of key terms and definitions.

2.1 Measure Coverage

Table 1 shows the patient population for the Breast Cancer Screening measure testing. It consists of Medicare beneficiaries enrolled in Medicare Parts A and B who meet all inclusion and exclusion criteria as specified by the measure.

Table 1: Beneficiary Demographics

Metric	Value
Number of Beneficiaries	4,150,266
Mean Age	72.88
Part D Enrollment %	78.48%

Table 2 shows the characteristics of groups and clinicians who are attributed at least 10 episodes.

Table 2: Group and Clinician Characteristics

Metric	Group		Clinician	
	Count	%	Count	%
Count	2,369	100%	17,464	100%
Number of Episodes Attributed	-	-	-	-
10-19 Episodes	65	2.74%	2,271	13.00%
20-39 Episodes	89	3.76%	2,414	13.82%
40-59 Episodes	75	3.17%	1,417	8.11%
60-79 Episodes	72	3.04%	1,042	5.97%
80-99 Episodes	38	1.60%	917	5.25%
100-199 Episodes	212	8.95%	2,907	16.65%
200-299 Episodes	171	7.22%	1,838	10.52%
300+ Episodes	1,647	69.52%	4,658	26.67%
Census Region	-	-	-	-
Northeast	451	19.04%	3,332	19.08%
Midwest	517	21.82%	4,176	23.91%
South	875	36.94%	6,515	37.31%
West	484	20.43%	3,368	19.29%

Metric	Group		Clinician	
	Count	%	Count	%
Unknown	42	1.77%	73	0.42%

2.2 Frequently Attributed Specialties

Table 3 shows the top 10 attributed specialties for this measure determined by the official CMS specialty designations listed on Medicare claims, using a 10-episode testing volume threshold. The most frequently attributed specialties reflect the intent of the measure to capture costs related to breast cancer screenings, including diagnostic radiology. These clinicians are also consistent with input provided by stakeholders, including patient and family partners (PFPs), during the measure development process.

Table 3: Count of the Top 10 Attributed Specialties

Specialty	Number of Clinicians Attributed
Diagnostic Radiology	13,723
Obstetrics/Gynecology	1,323
Family Practice	1,082
Internal Medicine	669
Nurse Practitioner	207
Interventional Radiology	175
Physician Assistant	108
General Surgery	34
Emergency Medicine	23
Hematology/Oncology	16

2.3 Reliability

Reliability evaluates a measure's ability to consistently differentiate the performance of one clinician from another. The signal-to-noise ratio is used to estimate reliability, which indicates how much of the variation in the measure score is explained by differences among clinicians' performance (i.e., signal) instead of differences within each clinician's performance (i.e., noise). Specifically, noise is the variation from one episode to another during the performance period for a particular clinician.

Table 4 shows reliability metrics at various testing volume thresholds. While higher thresholds yield higher reliability results, it is at the cost of further reducing the number of clinicians and clinician groups eligible for the measure, which would reduce the potential impact of the measure. For the purposes of field testing, we used a 10-episode testing volume threshold (bolded in the table below). If the measure is implemented in the Merit-based Incentive Payment System (MIPS) in the future, CMS will establish a case minimum through notice-and-comment rulemaking.

Table 4: Sample Size, Mean Reliability, and Proportion of Groups and Clinicians above Moderate Reliability at Various Testing Volume Thresholds

Testing Volume Threshold	Group			Clinician		
	Number of Groups	Mean Reliability	Percent Above 0.4	Number of Clinicians	Mean Reliability	Percent Above 0.4
10	2,369	0.98	100%	17,464	0.92	100%

Testing Volume Threshold	Group			Clinician		
	Number of Groups	Mean Reliability	Percent Above 0.4	Number of Clinicians	Mean Reliability	Percent Above 0.4
20	2,304	0.98	100%	15,193	0.95	100%
30	2,256	0.99	100%	13,791	0.96	100%

At the testing volume of 10 episodes, the mean reliability for the Breast Cancer Screening cost measure is high, specifically 0.98 at the group level and 0.92 at the clinician level (Table 4). 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.^{4,5} All groups and clinicians meet or exceed the moderate reliability threshold of 0.4 at the 10-episode testing volume threshold.

2.4 Validity

Validity is a criterion that evaluates whether the cost measure is able to quantify the construct that it aims to measure. Cost measures are designed to reflect the cost directly related to clinicians’ treatment choices, as well as the cost of adverse outcomes as a result of care.

Cost measure validity can be demonstrated in a conceptual model (Section 2.4.1). Validity can also be evaluated empirically by examining whether measure performance aligns with expectations from the conceptual model. That is, analyses can show to what extent adverse outcomes and treatment choices contribute to measure performance (Section 2.4.2).

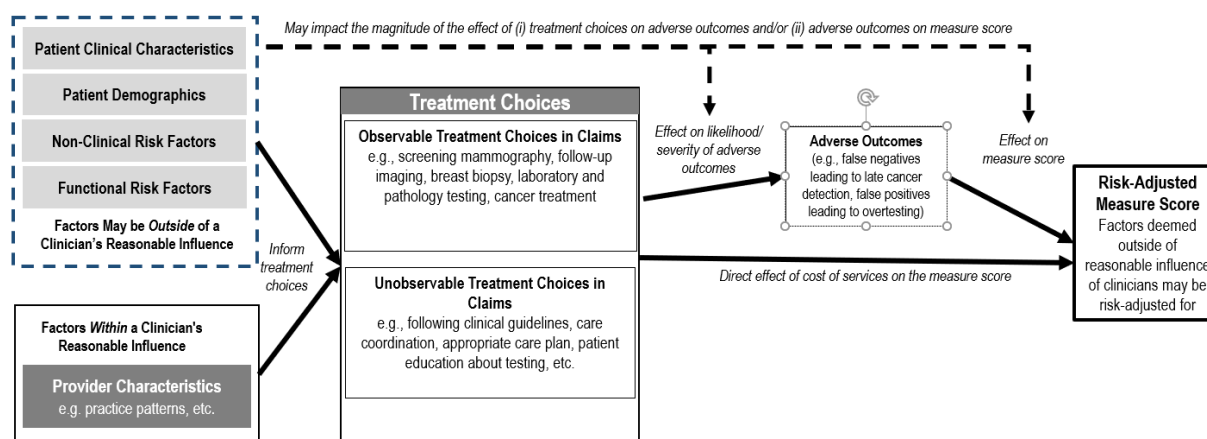
2.4.1 Conceptual Model

The measure conceptual model creates a comprehensive picture of the factors in and outside of a clinician’s control and how these factors interact with treatment choices, adverse outcomes, and the final measure score. Factors outside of a clinician’s reasonable influence include patient characteristics, patient demographics, and preexisting risk factors. These factors, combined with provider characteristics, impact the type, magnitude, and severity of treatment choices and adverse outcomes. The conceptual model shown in Figure 1 illustrates this relationship as well as how treatment choices and adverse outcomes impact the cost of services and affect the measure score.

⁴ Mathematica, Inc., “Memorandum: Reporting Period and Reliability of AHRQ, CMS 30-Day and HAC Quality Measures – Revised,” http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/HVBP_Measure_Reliability-.pdf.

⁵ CMS, “Medicare Program; CY 2022 Payment Policies Under the Physician Fee Schedule and Other Changes to Part B Payment Policies; Medicare Shared Savings Program Requirements; Provider Enrollment Regulation Updates; and Provider and Supplier Prepayment and Post-Payment Medical Review Requirements,” [86 FR 64996-66031](https://www.federalregister.gov/documents/2021/01/26/2021-01846).

Figure 1: Conceptual Model of the Relationship between Treatment Choices and the Measure Score



The Breast Cancer Screening cost measure assigns services that are clinically related to breast cancer screenings. Treatment choices observable in claims include screening mammograms, follow-up imaging (e.g., diagnostic ultrasounds, magnetic resonance imaging [MRI]), and breast biopsy, to name a few. Generally, unobservable treatment choices include following clinical guidelines, efficient care coordination, and appropriate care plans. These choices can affect the cost efficiency of care related to screening mammography and the likelihood and severity of adverse outcomes.

Adverse outcomes and inefficient care related to breast cancer screening can be caused by false-negative screening results (i.e., when cancer is missed during the screening mammography) resulting in late detection of breast cancer, and false-positive screening results (i.e., when cancer is incorrectly detected during the screening mammography) resulting in potentially unnecessary or inefficient follow-up diagnostics. Late detection often leads to more advanced staged cancers requiring high-cost diagnostic and treatment services; for this measure, late detection episodes are those in which cancer is detected between 9-12 months from the screening mammogram that triggers the episode. For more information on service assignment, please see the Draft Cost Measure Methodology.⁶

The measure intends to incentivize clinicians and groups who provide cost-efficient screening mammography, without disincentivizing appropriate use of follow-up diagnostics and cancer detection. To achieve this, the measure first assigns the costs of basic diagnostic services and certain complications to all episodes, but does not assign advanced diagnostic or oncology treatment costs unless late cancer detection occurs, as shown in Table 5. In other words, the measure is designed so that episodes with no cancer detection or early cancer detection will on average have better cost performance comparing to late cancer detection, while still allowing for assessment of cost efficiency. Additionally, for late detection episodes, oncology treatment costs are assigned using a fixed cost representative of oncology treatment costs, so that attributed groups and clinicians are not being held responsible for variation in treatment costs outside their reasonable influence (i.e., due to factors such as treatment complexity and tumor characteristics). The measure also risk adjusts for other factors that may fall outside a clinician's

⁶ The Draft Measure Methodology will be available on the CMS Cost Measures Information page once field testing begins: <https://www.cms.gov/medicare/quality/value-based-programs/cost-measures/current>.

reasonable influence on adverse outcomes (e.g., patient-level factors such as age, comorbidities, and disability status).

Table 5: Service Assignment Framework

Service Category	Assigned Clinically Related Services	Assigned to Which Episodes?
Basic Diagnostic Services	Mammography; diagnostic ultrasound; breast biopsy; MRI; evaluation and management (E/M) services (encounter for screening mammogram)	All
Emergency Department Services	Emergency department visits; critical care services	All
Advanced Diagnostic Services	Laboratory (chemistry and hematology); pathology; computed tomography (CT) scan; radioisotope scan and function studies	Only Late Cancer Detection
Treatment Services (assigned as a fixed cost)	E/M services (with breast cancer diagnosis); breast biopsy, local excision, and other breast procedures; mastectomy; lumpectomy, quadrantectomy of breast; cancer chemotherapy; anesthesia; non-hospital based care; CT scan for radiation therapy; therapeutic radiology; therapeutic procedures (skin and breast, female organs); ancillary services; medications (injections, infusions, etc.); durable medical equipment and supplies; hospitalizations (malignant breast disorders; septicemia or severe sepsis; complications of treatment)	Only Late Cancer Detection

2.4.2 Empirical Validity Analyses

To empirically assess validity, we assessed whether cost performance is better or worse for groups and clinicians based on cancer detection and timing.

Table 6 shows the distribution of mean observed-to-expected cost ratios (O/E ratios)—the ratio of actual costs to costs predicted by risk adjustment—across attributed episodes for groups and clinicians. A provider's mean O/E ratio is calculated for episodes by measure sub-group (cancer detection versus no cancer detection) and by timing of cancer detection, further distinguished as early versus late cancer detection. This breakdown is used to evaluate whether the measure incentivizes early cancer detection.

Table 6: Distribution of Mean O/E for Groups' and Clinicians' Episodes

Provider Level	Measure Characteristic	Provider Count	Distribution of Mean O/E Ratios					
			Mean	Percentile				
				10 th	25 th	50 th	75 th	90 th
Group	All Episodes	2,369	0.98	0.79	0.92	0.98	1.05	1.14
	No Breast Cancer Detection	2,369	0.98	0.79	0.91	0.98	1.05	1.14
	Breast Cancer Detection	2,023	0.99	0.53	0.76	0.99	1.21	1.41
	Early Cancer Detection	1,995	0.89	0.45	0.66	0.90	1.11	1.30
	Late Cancer Detection	1,108	1.93	1.01	1.33	1.85	2.38	2.96
Clinician	All Episodes	17,464	0.99	0.85	0.91	0.98	1.06	1.15
	No Breast Cancer Detection	17,464	0.99	0.84	0.92	0.98	1.06	1.15
	Breast Cancer Detection	10,409	0.99	0.39	0.66	0.96	1.25	1.57
	Early Cancer Detection	10,081	0.91	0.35	0.59	0.90	1.17	1.45
	Late Cancer Detection	2,593	1.89	0.96	1.08	1.75	2.45	3.23

Results show that groups and clinicians have better cost performance for episodes with early cancer detection compared to all other episodes. Groups' and clinicians' mean O/E ratios for episodes with an early cancer detection are on average 0.89 and 0.91, respectively. Results also show that providers have worse cost performance for episodes with a late cancer detection. Groups' and clinicians' mean O/E ratios for episodes with a late cancer detection are on average 1.93 and 1.89, respectively, which is more than 2 times higher than early detection episodes. This demonstrates that for episodes with a cancer detection, adverse outcomes (i.e., late cancer detection) affect the measure as intended.

2.5 Performance Gap

Table 7 shows the distribution of the measure score for clinicians and clinician groups meeting the testing volume threshold. These results align with expectations based on our review of the literature and demonstrate that there is a performance gap in cost measure performance between the most and least efficient groups and clinicians. Appendix B includes histograms for the distribution of group and clinician scores.

Table 7: Distribution of the Measure Score

Metric	Group	Clinician
Mean Score	\$221	\$223
Score Interquartile Range (IQR)	\$31	\$32
Standard Deviation	\$35	\$34
Coefficient of Variation	0.16	0.15
Score Percentile	Group	Clinician
10 th	\$179	\$191
25 th	\$207	\$206
50 th	\$222	\$222
75 th	\$237	\$238
90 th	\$258	\$259

Some variation is observed in the measure as indicated by the distribution of measure scores at both reporting levels. The measure score at the 90th percentile is 44% greater than the 10th

percentile (\$258 vs. \$179) at the group level, and 36% greater than the 10th percentile (\$259 vs. \$191) at the clinician level. Additionally, the interquartile range and standard deviation are both greater than \$30, showing some variation in measure scores. The results suggest that there is an opportunity for improvement in performance across groups and clinicians.

2.6 Risk Adjustment and Stratification

As shown in the conceptual model (Figure 1), patient-level and clinician-level factors can influence the measure score, which is informed by both published external research and our own data analysis.^{7,8,9,10} The conceptual model includes risk factors that are either known by the literature or informed by the Clinical Expert Workgroup to be within or outside of the influence of the attributed clinician. Risk factors, including non-clinical factors, can both influence the treatment choices and impact the size of the effect of treatment choices by mitigating the risk of adverse outcomes and the cost of adverse outcomes.

A systematic approach then guides the decision of which factors to include in the risk adjustment model. First, we reviewed the literature to gather known risk factors and drivers of resource use. These factors are usually diagnoses; therefore, the first set of risk adjustors are commonly the Hierarchical Condition Categories. Then, we consulted our clinical expert panels on additional factors that are known to be associated with resource use. Together with our clinical expert panel, we reviewed the stratified results on episode cost across many different patient characteristics. We arrived at the final list of risk adjustors based on those discussions and consensus among the clinical experts. Additionally, during our testing phases, we also follow a structured and systematic approach to decide whether eligibility for dual Medicare and Medicaid enrollment should be risk-adjusted for, which is further described in Section 2.7.

2.6.1 Discrimination

Discrimination is a statistical criterion that evaluates the measure's ability to distinguish high-cost episodes from low-cost episodes, or the ability to explain the variance in cost of individual episodes. The amount of variance explained is estimated by the R-squared metric with the range between 0 and 1. The R-squared value for the measure is 0.61, and remains the same after adjusting for the model's complexity based on the number of risk adjustors used. In other words, 61% of the variation in the actual observed cost of episodes is explained by the risk adjustment model and sub-group stratification.

The remaining unexplained variance is due to variation in factors that are not adjusted for by the measure, such as the clinician's performance. The objective of a cost measure is to evaluate and differentiate the performance of clinicians. Therefore, achieving high explained variance is not essential because not all of the variation in cost of care should be adjusted. In collaboration with the experts from our clinical workgroup, this measure only adjusts for factors that are deemed to be outside of the influence of clinicians. Please see the Draft Cost Measure Methodology for more information on the full list of risk adjustors and sub-groups.

⁷Assistant Secretary of Health and Human Services for Planning and Evaluation. Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. Washington, D.C. December 2016.

⁸Chen LM, Epstein AM, Orav EJ, Filice CE, Samson LW, Joynt Maddox KE. Association of Practice-Level Social and Medical Risk With Performance in the Medicare Physician Value-Based Payment Modifier Program. JAMA. 2017;318(5):453-461

⁹Medicare Payment Advisory Commission. Beneficiaries Dually Eligible for Medicare and Medicaid. 2018; <https://www.macpac.gov/publication/data-book-beneficiaries-dually-eligible-for-medicare-and-medicaid-3/>.

¹⁰ Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. 2020. <https://aspe.hhs.gov/social-risk-factors-and-medicare-value-based-purchasing-programs>

2.6.2 Calibration

Calibration evaluates the consistency of the measure in estimating episode cost across the full range of resource use patterns in the population. Table 8 assesses calibration by calculating the average predictive ratio for episodes partitioned into deciles based on expected episode cost. The predictive ratio is calculated using the formula of average expected cost divided by average observed cost for all episodes in each decile. A well-calibrated measure should have predictive ratios close to 1.00 across all deciles. In other words, such results show that the measure is consistent because it does not under- or over-predict cost throughout the range of resource use patterns in the population.

Table 8: Predictive Ratio by Decile of Predicted Episode Cost

Decile	Average Predictive Ratio
All	1.00
Decile 1	1.00
Decile 2	1.00
Decile 3	1.00
Decile 4	1.01
Decile 5	1.00
Decile 6	1.00
Decile 7	1.00
Decile 8	1.00
Decile 9	1.00
Decile 10	0.99

Table 8 demonstrates that the measure's risk adjustment model is consistent. Across all deciles, the average predictive ratios are approximately 1.00, and range between 0.99 and 1.0. Overall, the risk adjustment model does not over- or under-predict cost across the full range of resource use patterns in the population.

2.7 Dual Status and Provider Location Analysis

Beyond clinical characteristics of patients, the cost of care may be influenced by non-clinical factors such as location or coverage eligibility. Beneficiaries located in rural versus urban areas may face different barriers to accessing care. Additionally, beneficiaries eligible for dual Medicare and Medicaid enrollment status have been historically shown as a more vulnerable population, more likely to experience poor outcomes. At the program level, MIPS adjusts for these factors using the MIPS Complex Patient Bonus to ensure clinicians or groups treating more complex patients are not disadvantaged.¹¹ At the measure-level, testing helps to navigate the tension between ensuring fairness for clinicians treating higher shares of vulnerable patients and the possibility of masking poor performance and perpetuating disparity if clinicians are held to different standards.

Table 9 outlines the advantages and disadvantages of using dual status and provider claim location as indicators of non-clinical care factors.

¹¹ Quality Payment Program "COVID-19 Response," <https://qpp-cm-prod-content.s3.amazonaws.com/uploads/966/QPP%20COVID-19%20Response%20Fact%20Sheet.pdf>

Table 9: Tradeoffs of Dual Status and Claim Location Indicators

Variable	Advantages	Disadvantages	Used in Testing
Dual Medicare and Medicaid enrollment status	<ul style="list-style-type: none"> Available for all beneficiaries Most powerful predictor of poor outcomes¹² 	<ul style="list-style-type: none"> Variation in Medicaid eligibility across states 	Yes
Provider Location (i.e., urban or rural)	<ul style="list-style-type: none"> Urban or rural provider location can reflect potential differences in patient access or barriers to care 	<ul style="list-style-type: none"> Reflects provider location, not beneficiary residence, which may differ 	Yes

Table 10 presents the distribution of measure performance for groups and clinicians by provider location. Provider location (urban or rural) is determined by the Part B claim billing zip code capturing the highest total annual standardized cost from the 2024 study period or any year within a 2-year lookback window. The selected zip code per provider is then mapped to its zip locality 'rural' designation using the most recent quarter of mapping data available for the 2024 study period or any year within a 2 year lookback window. If the average urban and rural providers' scores are close to each other and to the national average provider score, it would indicate that the risk adjustment model adequately accounts for geographic differences in cost patterns.

Table 10: Associations of Provider Location for Groups and Clinicians

Provider Level	Provider Location	% of All Providers	Mean Provider Score	Distribution of Provider Score (Percentiles)				
				P10	P25	P50	P75	P90
Group	All Providers	100%	\$221	\$179	\$207	\$222	\$237	\$258
	Urban	77.42%	\$223	\$184	\$207	\$222	\$238	\$258
	Rural	22.58%	\$217	\$169	\$204	\$221	\$236	\$256
Clinician	All Providers	100%	\$223	\$191	\$206	\$222	\$238	\$259
	Urban	85.80%	\$223	\$192	\$207	\$222	\$239	\$259
	Rural	14.20%	\$220	\$180	\$204	\$220	\$237	\$256

Performance is similar for urban and rural providers at both provider levels. Most providers are located in an urban area, and have a mean provider score of \$223 at the group and clinician levels. Rural providers have a mean provider score of \$217 and \$220 at the group and clinician levels, respectively. The distribution of mean provider scores is also similar for urban and rural providers at both reporting levels. These results suggest that the risk adjustment model adequately accounts for geographic differences in cost patterns.

The subsequent analyses focus on dual status as the main proxy indicator of non-clinical factors for risk adjustment. Dual status indicates that a beneficiary is simultaneously enrolled in both Medicare and Medicaid. For testing purposes, a beneficiary will only be flagged as having dual

¹² Refer to footnote 4.

status if they are indicated as such in the Common Medicare Environment throughout the entirety of the episode window. To determine whether it's appropriate to risk adjust for dual status eligibility, the following criteria are considered:

- Association (Table 11): Is dual status associated with measure performance?
- Source (Table 11): Is the association patient-level or clinician-level?
- Complexity versus Quality (Tables 12, 13): Is the patient need, not poor quality, driving differences?
- Impact (Table 14): How would dual status adjustment affect measure performance rankings?

Table 11: Coefficient of Patient-level Dual Status under Different Models

Provider Level	Sub-group Risk Model	% of All Episodes	Coefficient of Patient-level Dual Status (P-value)		
			Base Model + Patient-level Dual Status	Base Model + Patient-level Dual Status + Clinician's Dual Share	Base Model + Patient-level Dual Status + Clinician's Fixed Effect
Group	No Breast Cancer Detection	99.08%	-3.17 (p < 0.0001)	-2.03 (p < 0.0001)	-2.11 (p < 0.0001)
	Breast Cancer Detection	0.92%	-94.41 (p = 0.02)	-98.36 (p = 0.02)	-90.71 (p = 0.03)
Clinician	No Breast Cancer Detection	99.08%	-3.28 (p < 0.0001)	-1.97 (p < 0.0001)	-2.15 (p < 0.0001)
	Breast Cancer Detection	0.92%	-90.78 (p = 0.03)	-109.84 (p = 0.01)	-84.34 (p = 0.08)

Table 12: Mean O/E Ratio Stratified by Dual Share and Patient's Dual Status

Share of Episodes with Dual Status	Group			Clinician		
	All Episodes	Dual Episodes	Non-Dual Episodes	All Episodes	Dual Episodes	Non-Dual Episodes
All	0.98	0.96	0.98	0.99	0.97	0.99
0-20%	1.00	0.99	1.00	0.99	0.99	0.99
21-40%	0.99	0.97	0.99	1.00	0.98	1.00
41-60%	1.00	0.98	1.00	0.99	0.98	0.99
61-80%	0.97	0.96	0.98	0.99	0.98	0.99
81-100%	0.94	0.93	0.94	0.97	0.96	0.98

Table 13: Proportions of Groups and Clinicians Who Perform Significantly Worse, Equally Well, or Significantly Better on Their Dual Episodes than Non-Dual Episodes

Provider Level	Significantly Worse	Equally Well	Significantly Better
Group	4.03%	84.79%	11.18%
Clinician	3.52%	93.03%	3.45%

Table 14: Group and Clinician Performance Shift Measured by the Change in Measure Score After Adding a Patient-level Dual Status Risk Adjustor

Provider Level	Proportions of Groups and Clinicians Affected at Various Levels of Performance Shift	
	Ranking Shift by 1% or more	Ranking Shift by 5% or more
Group	18.32%	0.13%
Clinician	16.78%	0.10%

Table 11 shows there is a statistically significant negative association between the patient's Medicaid enrollment status and episode cost for both clinicians and groups in the no breast cancer detection sub-group; that is, clinicians and groups are expected to spend less on episodes in which the patient is dually enrolled in Medicaid. However, performance for episodes with and without dual enrollment remain relatively stable as clinician's shares of episodes with dual enrollment increases (Table 12). Table 13 demonstrates that more than 80% of groups and more than 90% of clinicians perform equally well for episodes with and without dual enrollment. Some clinicians (11.18%) perform significantly better on episodes with dual enrollment at the group level (Table 13). Moreover, risk adjusting for dual enrollment status appears to change measure performance for some clinicians, but few clinician's ranks shift by 5% or more (Table 14). These results suggest that clinicians are able to mitigate many effects of non-clinical factors, or that the current risk adjustment model sufficiently adjusts for non-clinical factors.

2.8 Impact of Exclusions

Table 15 displays descriptive statistics of all episodes meeting the measure's triggering logic, excluded episodes, and final reportable episodes at both group and clinician levels. These exclusion criteria ensure that the reportable episode populations are more homogenous and comparable than all episodes meeting triggering logic. It is worth noting that only the observed cost is shown, which has not been risk adjusted for using our risk adjustment model. Therefore, the differences in cost may appear much smaller after risk adjustment than as-is.

Table 15: Cost Statistics for Measure Exclusions

Exclusion Criteria	Episodes		Mean	Observed Episode Cost Percentile				
	Count	Percent of All Episodes Meeting Trigger Logic		10 th	25 th	50 th	75 th	90 th
All Episodes Meeting Triggering Logic	4,598,257	100.00%	\$243	\$173	\$180	\$180	\$216	\$336
Beneficiary Death in Episode	37,895	0.82%	\$235	\$128	\$180	\$180	\$216	\$245
Outlier Cases	83,219	1.81%	\$602	\$36	\$64	\$567	\$807	\$1,547
Not in OP, IP, or ASC Setting	26,657	0.58%	\$228	\$128	\$180	\$180	\$180	\$306
Group does not Meet Testing Volume Threshold	5,114	0.11%	\$375	\$116	\$128	\$180	\$308	\$1,117
Clinician does not Meet Testing Volume Threshold	25,695	0.56%	\$283	\$116	\$144	\$180	\$216	\$612
Age Under 40	815	0.02%	\$254	\$128	\$180	\$180	\$216	\$349
History of Breast Cancer	376,616	8.19%	\$354	\$180	\$180	\$180	\$314	\$536
Not Female	112	0.00%	\$209	\$64	\$93	\$180	\$180	\$335
Reportable Episodes (if all clinicians reported as Group at the Testing Volume Threshold)	4,080,337	88.74%	\$226	\$173	\$180	\$180	\$216	\$308
Reportable Episodes (if all clinicians reported as Clinician at the Testing Volume Threshold)	4,064,587	88.39%	\$226	\$174	\$180	\$180	\$216	\$308

The statistical results show that the reportable episode populations are more homogenous and comparable than all episodes meeting triggering logic since excluding these episodes reduce variation in the distribution of observed costs. For instance, the observed cost of all episodes meeting trigger logic had a mean observed cost of \$243 and ranged from \$173 in the 10th percentile to \$336 in the 90th percentile. Meanwhile, the reportable episodes at the group level had a mean observed cost of \$226 with a smaller distribution ranging from \$173 in the 10th

percentile to \$308 in the 90th percentile. Similarly, the reportable episodes at the clinician level had a mean observed cost of \$226 ranging from \$174 in the 10th percentile to \$308 at the 90th percentile.

The excluded cohorts generally have higher mean observed costs compared to all episodes meeting triggering logic. Episodes with a beneficiary death; episodes that do not take place in an outpatient, inpatient, or ambulatory surgical center (ASC); and episodes without a female patient have lower mean observed costs compared to all episodes meeting triggering logic, but make up less than 1% of all triggered episodes. The largest exclusions come from patients with a history of breast cancer, which is around 8% of all episodes meeting triggering logic. Overall, these findings support the exclusion of these episodes to ensure a comparable patient cohort that will yield a clinically coherent measure and meaningful information to attributed clinicians.

Appendix A. Glossary

Table A1: Key Terms and Definitions

Term	Definition
Attributed Clinician/ Group	The clinician or group determined to be responsible for an episode based on the measure's attribution methodology. Please see the Draft Cost Measure Methodology for a description of the attribution methodology and the Draft Measure Codes List for the list of codes in measure attribution. ¹³
Calibration	A statistical property that evaluates whether a risk adjustment model produces accurate cost predictions across the full range of patient complexity. A well-calibrated model has predictive ratios close to 1.0 across all risk levels.
Coefficient of Variation	A measure of relative variability, calculated as the standard deviation divided by the mean. Higher values indicate greater variability in measure scores.
Discrimination	A statistical property that evaluates how well a risk adjustment model distinguishes between high-cost and low-cost episodes. Measured by R-squared.
Dual Eligible (Dual Status)	Beneficiaries who qualify for both Medicare and Medicaid. Dual-eligible beneficiaries often have greater health needs and may face additional barriers to care.
Episode	A defined period during which clinically related services are delivered to a patient for a specific condition or procedure. Episode-based cost measures group these services together to assess resource use.
Expected Cost	The cost predicted by the risk adjustment model for an episode, based on patient characteristics, risk factors, and other variables included in the model. Expected cost serves as the benchmark against which observed cost is compared to calculate the O/E ratio.
Hierarchical Condition Categories (HCCs)	A risk adjustment methodology that uses diagnosis codes to predict healthcare costs. HCCs group clinically related conditions and assign risk scores based on expected resource use.
Interquartile Range (IQR)	The difference between the 75th percentile and 25th percentile of a distribution. The IQR represents the range containing the middle 50% of values.
Merit-based Incentive Payment System (MIPS)	A program established by the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) that adjusts Medicare payments to clinicians based on performance in quality, cost, improvement activities, and Promoting Interoperability.
National Provider Identifier (NPI)	A unique 10-digit identification number assigned to healthcare providers. Used in combination with TIN to identify individual clinicians (TIN-NPI).
Observed Cost	The actual, payment standardized, Medicare payments made for services during an episode, before any risk adjustment.
Observed-to-Expected Cost Ratio (O/E Ratio)	The ratio of actual (observed) episode costs to the costs predicted (expected) by the risk adjustment model. An O/E ratio above 1.0 indicates higher-than-expected costs; below 1.0 indicates lower-than-expected costs.
Predictive Ratio	The ratio of average expected cost to average observed cost for a group of episodes. Used to assess calibration. Values close to 1.0 indicate good calibration.

¹³ These documents will be available on the CMS Cost Measures Information page once field testing begins: <https://www.cms.gov/medicare/quality/value-based-programs/cost-measures/current>.

Term	Definition
Reliability	A measure of how consistently a measure differentiates performance between clinicians. Expressed as a value between 0 and 1, where higher values indicate that observed differences in scores more likely reflect true differences in performance rather than random variation.
Risk Adjustment	A statistical method that accounts for differences in patient characteristics (such as age, health status, and comorbidities) that may affect costs but are outside the clinician's control.
Risk-Adjusted Cost	Episode cost after applying the risk adjustment model to account for patient characteristics and other factors.
R-squared (R^2)	A statistical measure indicating the proportion of variance in episode costs explained by the risk adjustment model. Values range from 0 to 1, with higher values indicating the model explains more of the variation in costs.
Signal-to-Noise Ratio	A measure used to estimate reliability. "Signal" represents true differences in clinician performance; "noise" represents random variation within a clinician's episodes.
Stratification	The division of episodes into sub-groups based on patient or clinical characteristics to improve the accuracy of expected cost calculations.
Tax Identification Number (TIN)	A number used to identify a business entity (such as a group practice) for tax purposes. Used to identify clinician groups in measure reporting.
TIN-NPI	A combination of Tax Identification Number and National Provider Identifier used to identify an individual clinician practicing within a specific group.
Validity	The extent to which a measure accurately captures the concept it is intended to measure. For cost measures, validity assesses whether the measure reflects costs related to treatment choices and outcomes.

Appendix B. Distributions of Measure Score

Figure B1: Distribution of Measure Score - Group

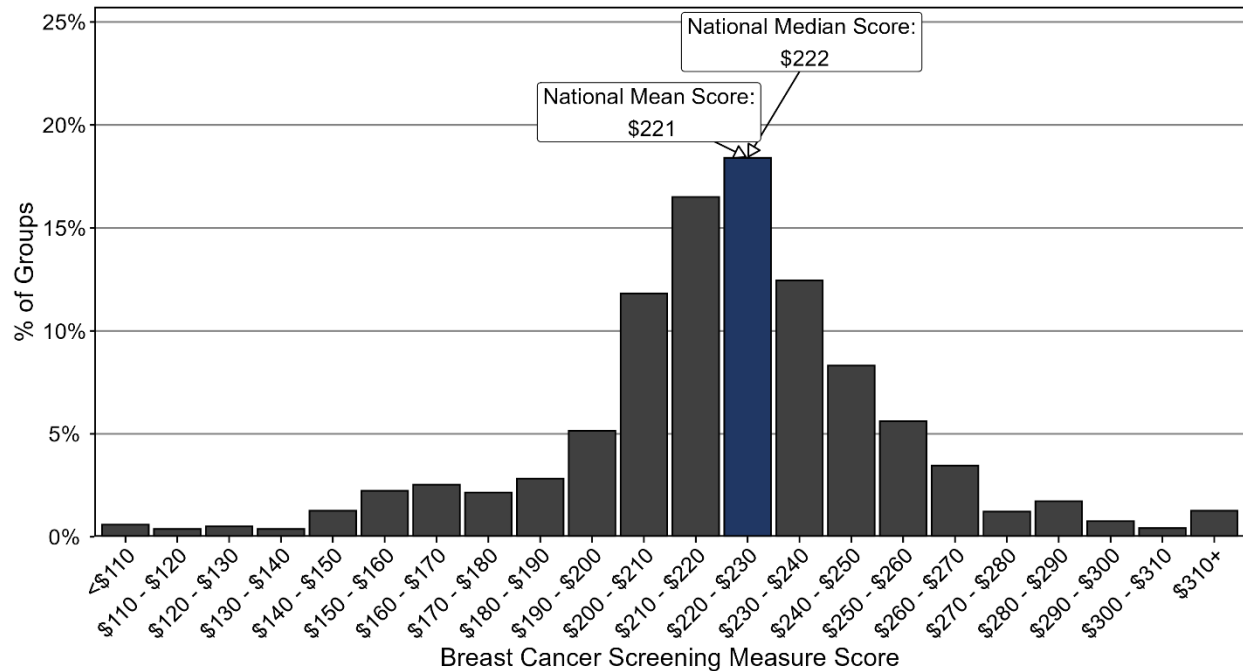


Figure B2: Distribution of Measure Score – Clinician

