OP-39 Re-specification Technical Expert Panel (TEP) Meeting Summary

Development, Reevaluation, and Implementation of Outpatient Outcome/Efficiency Measures
OP-39 Re-specification TEP Meeting: July 10, 2025
August 2025

Contents

1. Project Overview

The Centers for Medicare & Medicaid Services (CMS) has contracted with Acumen, LLC, its non-profit partner firm The SPHERE Institute, and subcontractors (hereafter jointly referred to as "Acumen") to develop, reevaluate, and implement CMS outpatient outcome and efficiency measures for CMS public reporting and payment programs. As part of its measure development and re-specification process, Acumen convenes groups of stakeholders who contribute direction and thoughtful input to the measure developer during measure development and maintenance.

Acumen solicited nominations for a Technical Expert Panel (TEP) composed of stakeholders and experts to contribute direction and thoughtful input to re-specify the Breast Cancer Screening Recall Rate measure (OP-39) that is currently in use in the Hospital Outpatient Quality Reporting (Hospital OQR) program. The nomination period was from May 14, 2025 to June 17, 2025. The OP-39 measure was submitted for Consensus-Based Entity (CBE) endorsement during the Fall 2023 cycle but did not receive endorsement. The purpose of this TEP is to provide input for the re-specification of the OP-39 measure in order to address the limitations and weaknesses discussed during the endorsement cycle.

2. OP-39 Re-specification TEP Meeting: July 10, 2025

This meeting summary document outlines the purpose, discussion, and recommendations from the OP-39 Re-specification TEP meeting. Section 2.1 provides an overview of the meeting goals and process. Section 2.2 summarizes the discussion and recommendations from the TEP.

2.1 Overview

The goal of the TEP was to provide input on the following topics that could assist with potential re-specification of the OP-39 measure:

(i) Discuss measure scoring methodologies that better reward strong performance in the hospital outpatient setting
(ii) Review approaches to account for differences in patient sub-populations
(iii) Discuss possible companion measures and other approaches to provide hospitals with additional information they can use to improve performance

The meeting was held online via webinar and attended by 10 of the 11 TEP members. The webinar was facilitated by Acumen moderator, Heather Litvinoff. The TEP Chair was David Seidenwurm, who also facilitated meeting discussions. Four Person and Family Engagement (PFE) TEP members (i.e., members

with relevant experience as a patient, family/caregiver representative) attended the meeting, which included Heather Guidone, Karen Fernandes, and Rosie Bartel.[1] The OP-39 Re-specification TEP Membership List contains the full list of members, including names, professional roles, employers, and clinical specialties, where applicable.[2]

Prior to the meeting, TEP members were provided with information and materials to inform the discussions. This included the meeting agenda and slide deck with embedded empirical results from testing analyses.

After the meeting, TEP members were sent a recording of the webinar and were polled on their preferences to ensure the measure re-specification recommendations are based on formal, well-documented input. Based on similar meeting discussion practices, the threshold for support was >60% consensus among poll responses. This document summarizes the TEP members' input from both the discussion as well as the poll.

This meeting was convened by Acumen as part of the measure re-specification process to gather expert input; as such, these are preliminary discussions and materials, which don't represent any final decisions about the measure specifications.

2.2 Summary of Sessions and Discussion

This section is organized based on meeting sessions and describes TEP member discussions and recommendations. During the first portion of the meeting, Acumen facilitated introductions for TEP members, provided project background, and discussed the goals for the meeting. Next, Acumen described the OP-39 measure, provided context on the measure's development and implementation, and described the current measure specifications. As part of this, Acumen outlined the feedback the measure received during the 2023 CBE endorsement process, which is provided below in Table 1 along with the proposed solutions that the TEP discussed during the meeting.

Table 1: 2023 CBE Endorsement Process Weakness Areas and Proposed Solutions

Weakness Area	Proposed Solution
1. Lack of Connection Between Acceptable Benchmark and Hospital Performance or Quality of Care	• Create a scoring methodology to rank hospitals based on their proximity to the acceptable benchmark range • Incorporate additional elements into OP-39 that indicate a quality of care
2. Lack of Risk Adjustment and Flexibility in the Acceptable Benchmark Range	• Apply exclusions to focus on a more homogenous patient population
3. Lack of Connection Between Measure Score and Actions Hospitals Could Take to Improve Performance	• Supplement future CBE submission with additional literature on interventions hospitals can implement to improve performance
4. Lack of Patient Engagement During Development	• Convene this OP-39 Re-specification TEP, including Person and Family Engagement (PFE) members

[1] One additional PFE TEP member attended the meeting, but they opted to have their name and information not included in official TEP materials such as this summary report.
[2] The OP-39 Re-specification TEP Membership List is available on the "CMS Updates to Established TEPs" Page (https://mmshub.cms.gov/get-involved/technical-expert-panel/updates).

During the meeting, the TEP Chair, David Seidenwurm, and TEP moderator, Heather Litvinoff, addressed other feedback and questions from TEP members on the OP-39 measure. For example, a member asked how and why the 45-day window was selected for the measure. Dr. Seidenwurm responded that this was based on historical consensus among previous TEPs as well as the most up-to-date data from that time. Specifically, he noted that the data from the previous TEPs showed a precipitous drop beyond 45 days for the duration of time between the index and follow-up studies.

The remaining sub-sections in this report describe TEP member discussions and recommendations on Weakness Areas 1 through 3 from Table 1 above: creating a sensible scoring methodology, addressing patient cohort heterogeneity concerns, and providing actionable information to hospitals, respectively. The final sub-section provides an overview of next steps for the measure re-specification process.

2.2.1 Creating a Sensible Scoring Methodology

Acumen discussed the first weakness area from the 2023 CBE endorsement process feedback, which was the lack of connection between acceptable benchmark and hospital performance or quality of care. This section was split into two sub-sections: (i) presenting options for a scoring methodology that ranks hospitals based on their proximity to the measure's acceptable benchmark range, and (ii) discussing the Positive Predictive Value (PPV) measures as options for incorporating additional information for the OP-39 measure to better capture quality of care.

First, Acumen discussed how the current measure scoring approach does not incorporate the target recall rate, which the American College of Radiology (ACR) recommends as between 5 and 12 percent. Rather, the OP-39 measure is currently configured to measure recall rates between 0 and 100 percent, with lower values generally indicating higher quality. However, as the ACR notes in its recommendation, there are potentially negative consequences for the rate being either too low or too high: values below 5 percent may indicate hospitals could be systematically missing cancer cases while values above 12 percent may indicate they may be recalling too many cases for follow-up, which increases cost, anxiety for patients, and radiation from screening.

The moderator addressed a TEP member's question by clarifying that the OP-39 measure does not penalize hospitals for their performance, as the Hospital OQR program currently just requires hospitals to submit their data for the measure.

The TEP reviewed historical data for public reporting (PR) years 2022 through 2025, showcasing how there has been a growing share of hospitals below 5 percent and a decreasing share above 12 percent. Dr. Seidenwurm remarked on how these results showcase radiologists have been improving over time in terms of recall rate, as the mean measure score fell from around 11 percent to 7 percent, which he noted as a more optimal "sweet spot" for recall rate based on previous research. One TEP member noted that this improvement likely stems from the implementation of the OP-39 measure as well as a broadening use of tomosynthesis.

Acumen introduced three dimensions to consider for an alternative scoring methodology that can rank hospitals based on their proximity to the acceptable range. These dimensions, along with 4 alternative scoring methodologies, are presented in Table 2.

Table 2: Baseline and Alternative Scoring Methodologies for the OP-39 Measure

Option	**Dimension 1:** Should performance within the acceptable range be rewarded?	**Dimension 2:** Should there be differentiation within the acceptable range?	**Dimension 3:** Should the impact of being outside the acceptable range by symmetrical?
Baseline	No	No	Not Applicable
A	Yes	No	Yes
B	Yes	Yes, Lower Recall is Better	Yes
C	Yes	Yes, Lower Recall is Better	No, Hospitals Receive **More Favorable Ranking** for Being **Below** the Acceptable Range than Above
D	Yes	No	No, Hospitals Receive **More Favorable Ranking** for Being **Below** the Acceptable Range than Above

Each of Options A through D address the underlying issue with the baseline scoring approach, by providing a more favorable ranking to hospitals within the acceptable range compared to those outside the range (Dimension 1). For hospitals within the acceptable range, Options A and D offer no differentiation across hospitals depending on their specific rate; however, Options B and C are configured to more favorably rank those that are on the lower end of the acceptable range (Dimension 2). This means a hospital with a rate of 5 percent will have a higher (i.e., better) rank than one with a rate of 12 percent for Options B and C. For hospitals outside the acceptable range, Options A and B offer symmetrical impact depending on whether they are below or above the acceptable range; however, Options C and D are configured to more favorably rank hospitals below the acceptable range relative to those above the range (Dimension 3). This means a hospital with a rate below 5 percent will always have a higher rank than one with a rate above 12 percent for Options C and D.

Acumen also presented data on how hospitals perform under the current (Baseline) approach and the 4 alternative scoring methodologies based on whether they are within or outside the cancer detection rate range recommend by the ACR (i.e., 2.5 or more cancers detected per 1,000 screenings). The Baseline approach currently yields a higher score for hospitals outside of the ACR range (i.e., for hospitals with less than 2.5 cancers detected per 1,000 screenings); however, all 4 alternative scoring methodologies provide more favorable ranking for hospitals that are within the ACR range.

A TEP member recommended not having a differential within the ACR range (Dimension 2) because the recall rate is affected by many factors (e.g., the subspecialty training or age of the breast imager, changing technologies with artificial intelligence and tomosynthesis), and several other members agreed. Other members felt that from a statistical standpoint, there may not be statistical significance and differentiating within the range would give false precision. TEP members also discussed worrisome incentives of driving lower recall rates, but they felt that if reviewed alongside other measures, incentives could be balanced. A member raised that a scoring methodology without a differential for hospitals within the ACR range may yield issues with the measure becoming topped-out. The PR 2025 data indicates that over 62 percent of hospitals are within the ACR range, meaning the majority would all receive a maxed out ranking of 100 in an alternative scoring methodology that offers no differential among hospitals within the ACR range.

One TEP member stated it was unclear whether rates below or above the ACR range for OP-39 should be considered a worse outcome, suggesting the measure may not want privilege the “below range” group over the “above range” group. This would support the alternative methodologies with symmetrical impact for hospitals on either end of the ACR range (i.e., Options A and B). Another member agreed that this question cannot be addressed without viewing the cancer detection rate alongside OP-39 performance. However, another member stated that the risk of missing cancer cases is

greater than the risk of recalling too many cases, which Dr. Seidenwurm indicated as in support of an asymmetrical approach for Dimension 3 (i.e., hospitals should receive more favorable ranking for being above the acceptable range compared to those below the range).

Next, Acumen provided some background on how clinical guidelines for breast imaging recommend considering cancer detection metrics simultaneously to provide valuable information to hospitals. In particular, the ACR Breast Imaging Reporting and Data System (BI-RADS)® Atlas, 5th edition breast imaging medical audit calls for the use of specific PPV measures. Measure descriptions for the 2 PPV measures are provided below in Table 3.

Table 3: Measure Descriptions for PPV Measures & Their ACR Target Ranges

Item	PPV-1	PPV-3
Name	**Breast Cancer Screening with an Eventual Breast Cancer Diagnosis**: Positive Predictive Value 1 (PPV-1)	**Use of Biopsy After Diagnostic Follow-up with an Eventual Breast Cancer Diagnosis**: Positive Predictive Value 3 (PPV-3)
Description	Share of **women 40 years of age and older** who had a positive screening mammogram that led to an eventual breast cancer diagnosis in the Medicare population. Positive screening mammograms are measured as cases with diagnostic follow-up for additional testing, including diagnostic digital breast tomosynthesis (DBT), diagnostic mammogram, ultrasound, and magnetic resonance imaging (MRI).	Share of **women 40 years of age and older** who had a **biopsy** from a **diagnostic follow-up** that resulted in an **eventual breast cancer diagnosis** in the Medicare population
Denominator	**Women 40 years of age and older** who received **diagnostic follow-up** (including diagnostic DBT, diagnostic mammogram, ultrasound, and MRI) within **45 days of a screening mammogram**	**Women 40 years of age and older** who received a **biopsy within 45 days of a diagnostic follow-up** (including diagnostic DBT, diagnostic mammogram, ultrasound, and MRI)
Numerator	**Women 40 years of age and older** who received: (i) services related to **breast cancer treatment,** or (ii) at least **two evaluation and management (E/M) services** on separate days, both with a diagnosis indicating **breast cancer**, **within 8 months of the screening mammogram**	**Women 40 years of age and older** who received: (i) services related to **breast cancer treatmen**t, or (ii) at least **two evaluation and management (E/M) services** on separate days, both with a diagnosis indicating **breast cancer**, **within 4 months of the diagnostic follow-up**
ACR Target Range	**3 – 8%**	**20 – 45%**

Acumen noted that a separate TEP was convened in 2023-4 to provide input for the development of a suite of breast cancer screening cost and quality measures intended for use in the Merit-based Incentive Payment System (MIPS) at the clinician and clinician group levels. This included a breast cancer screening recall rate measure, similar to the OP-39 measure, as well as the PPV-1 and PPV-3 measures described in the table above. The TEP reviewed the measure specifications for the PPV measures, including their ACR target ranges and PR 2025 data on their performance for outpatient hospitals.

The TEP discussed the possibility of combining the OP-39 measure with the PPV measures in a composite measure, and they reviewed similar data on the 4 alternative scoring methodologies for a composite measure approach. However, the PR 2025 data indicated a substantial decrease in the number of hospitals that would meet the case minimum for a composite measure approach (i.e., from 3,343 hospitals down to 860). This is largely driven by the comparatively small number of outpatient hospitals that meet a case minimum of 20 denominator cases for the PPV-3 measure (i.e., 866). One TEP member asked whether it may be more useful to use cancer detection rate (i.e., cancers detected per 1,000 screenings) instead of PPV-3. Dr. Seidenwurm noted that for cancer detection rate and the 2.5

threshold recommended by the ACR, this would likely yield a topped-out measure, meaning there isn't enough variation across measured entities. TEP members also raised that the ACR ranges or thresholds for recall rate, the PPVs, and cancer detection rate may need to be adjusted or revisited for this older Medicare population. For example, the cancer detection rate threshold may need to be higher than 2.5 for the Medicare population, as there are more cancers detected among Medicare beneficiaries relative to the general population. Dr. Seidenwurm asked the group whether the OP-39 measure, as well as the potential companion measures, ought to move forward with the ACR recommended ranges. It was mentioned that the ACR recommended ranges were developed for a general population, whereas the Medicare population is older any may need some adjustment. One member recommended remaining with the ACR range for OP-39, as it's widely known and embedded into systems audits and breast imaging practices, though the ACR range may change over time due to advances with artificial intelligence (AI). Although some members seemed to be open to changing the acceptable benchmark range from the ACR's current guidelines, without an alternative set of guidelines or recommendations to point to, the general consensus from the TEP was to maintain the current range.

One PFE TEP member recommended extending the time window for OP-39, which is currently 45 days between the index and follow-up studies, since some patients (e.g., those in rural areas) may not be able to receive the follow-up study within that timeframe. Another PFE TEP member noted that 45 days is a long time and suggested a shorter time window, like 30 days.

<u>Key Takeaways from Discussion and/or Polls for Creating a Sensible Scoring Methodology</u>:

- Members unanimously recommended to update the scoring methodology such that hospitals within the ACR acceptable range have more favorable ranking than hospitals outside the range.
- The TEP reached consensus to update the scoring methodology such that hospitals within the ACR acceptable range receive the same favorable ranking.
- The TEP didn't achieve consensus on whether the impact of being above or below the ACR acceptable range should be symmetrical.
- In terms of ranking Options A through D, over 90 percent of TEP members ranked Option A and over 63 percent ranked Option D as one of their top 2 alternative options; less than 37 percent of members ranked Options B and C as one of their top 2 alternative options.
- Members reached consensus that the PPV-1 and Cancers Detected per 1, 000 Screenings measures would add value to the OP-39 measure when optimally specified; they didn't reach consensus for the PPV-3 measure.
- In terms of ranking the three companion measures regarding the value they'd add to the OP-39 measure, over 81 percent of TEP members ranked PPV-1 and Cancers Detected per 1, 000 Screenings as one of their top 2 measures; less than 37 percent of members ranked PPV-3 as one of their top 2 measures.
- Most members preferred reporting the companion measures as distinct but complementary measures, rather than combining them into a composite. Key reasons included:
    - A composite measure approach would yield a significant reduction in the number of hospitals that would meet the case minimum, which would disproportionately exclude smaller, rural, or underserved hospitals from reporting and quality improvement.
    - Reporting measures separately enhances clarity and provides interested parties, including patients and clinicians, more actionable and understandable insights.
- The TEP did not form consensus on whether to update the window from the current 45 days.

2.2.2 *Addressing Patient Cohort Heterogeneity Concerns*

Acumen discussed the second weakness area from the 2023 CBE endorsement process feedback, which was the lack of risk adjustment and flexibility in the acceptable benchmark range. Dr. Seidenwurm noted that the OP-39 measure does not require risk adjustment, since high-risk sub-populations have traditionally exhibited recall rates comparable to the broader population. TEP members also received PR 2025 data that demonstrated that an outpatient hospital's share of dually eligible beneficiaries had minimal impact on measure score performance. Therefore, the focus for addressing this point of concern was to consider applying exclusions to the measure in order to focus on a more homogenous patient cohort. The OP-39 measure currently does not have any exclusions.

The TEP reviewed some background for the development of the clinician-level breast cancer screening cost and quality measures to consider for potential patient cohort alignment. Acumen noted that the following measure exclusions were recommended for this suite of breast cancer screening measures: (i) male beneficiaries, (ii) beneficiaries under the age of 40, and (iii) beneficiaries with a history of breast cancer; several additional sub-populations for potential exclusion were discussed, such as beneficiaries with genetic risk of breast cancer (BReast CAncer genes [BRCA] carrier status), dense breast tissue, and abnormal mammograms. Acumen noted how guidelines across organizations (e.g., ACR, American Cancer Society, American College of Obstetricians and Gynecologists) consistently cite age 40 for when women ought to begin receiving mammograms. The TEP reviewed and discussed the PR 2025 data indicating the size of these measure sub-populations, which is presented below in Table 4.

Table 4: PR 2025 Data on OP-39 Measure Sub-Populations of Interest

Sub-Population	Denominator Cases		Beneficiaries	
	#	%	#	%
All Denominator Cases	3,599,326	100.00%	3,597,827	100.00%
Less than 40 Years of Age	1,120	0.03%	1,120	0.03%
Male	360	0.01%	360	0.01%
History of Breast Cancer	316,186	8.78%	315,810	8.78%
History of Genetic Risk of Breast Cancer (BRCA Carrier Status)	8,626	0.24%	8,617	0.24%
Prior Presence of Dense Breast Tissue	118,140	3.28%	118,063	3.28%
History of Abnormal Mammogram	149,521	4.15%	149,420	4.15%

One PFE TEP member asked about men developing breast cancer, which Dr. Seidenwurm addressed by noting it's very rare and typically associated with BRCA carrier status; this member recommended keeping beneficiaries with BRCA carrier status as well as those exposed to diethylstilbestrol (DES) in the measure. Another member recommended excluding beneficiaries less than 40 years of age, males, and those with a history of breast cancer, suggesting the aim should be for a relatively homogenous population that can be easily identified. He also noted that we may want to also exclude beneficiaries with genetic risk markers, though it shouldn't be limited to only BRCA carrier status, since there are many more (e.g., Lynch Syndrome, Li-Fraumeni), and they are rare.

Key Takeaways from Discussion and/or Polls for Addressing Patient Cohort Heterogeneity Concerns:

- Members achieved consensus on excluding the following sub-populations for the OP-39 measure:
  - Beneficiaries less than 40 years of age
  - Male beneficiaries
  - Beneficiaries with a history of breast cancer
- The TEP reached consensus to not exclude beneficiaries with a history of genetic risk of breast cancer (BRCA carrier status).

2.2.3 *Providing Actionable Information to Hospitals*

Acumen discussed the third weakness area from the 2023 CBE endorsement process feedback, which was the lack of connection between measure score and actions hospitals could take to improve performance. The TEP discussed how the PPV measures, discussed above, will help to address this concern, providing hospitals with valuable information on cancer detection to pair alongside recall rate. However, beyond the value brought by incorporating the PPV measures, Acumen also presented several strategies that hospitals may explore to help improve their OP-39 measure performance based on a comprehensive scan of literature. These strategies are listed and described below in Table 5.

Table 5: Additional Strategies to Improve OP-39 Measure Performance

Strategy	Description
Radiologist Specialization	Ensure radiologists have formal training in breast imaging or over 5k mammograms annually
Double Reading	Implement consensus double reading, requiring agreement between 2 radiologists before recalling a patient
Staff + Physician Education	Train front-desk staff to encourage patients to bring prior exams + referring physicians on ensuring to specify between screening v. diagnostic mammography
Centralized Image Interpretation	Centralize the reading of mammograms by breast imaging specialists
Modality Assessment and Adjustment	Determine the value of other imaging modalities in improving recall rate (as well as PPV metrics); this may result in incorporating DBT into screening protocols to improve diagnostic accuracy
Regular Performance Audits	Conduct periodic reviews of recall cases + outcomes to identify patterns and areas for improvement; this may include establishing a system for timely follow-up after breast cancer screening (i.e., immediate referral for further testing of abnormal results)

The TEP discussed additional strategies and other important elements of information that can be explored in future testing to aid hospitals in having more actionable information to improve their OP-39 measure performance.

TEP members recommend adding the following to the list of strategies: (i) using AI for mammograms, (ii) measuring cancer detection rate alongside the OP-39 measure, and (iii) taking radiologist groups' performance into account, including their minimum volume requirements. A PFE TEP member agreed that using AI for breast imaging would be beneficial to ensure more beneficiaries can get more readings, but she noted that there should be guardrails (e.g., radiologist involvement or supervision for critical thinking components) to ensure the quality of the readings.

<u>Key Takeaways from Discussion and/or Polls for Providing Actionable Information to Hospitals</u>:

- Members offered several strategies to provide actionable information to hospitals, including the following:
    - Stratify OP-39 data by factors such as age, race/ethnicity, and socioeconomic status to identify disparities and guide targeted interventions
    - Consider facility-level characteristics (e.g., size, location, volume) to understand systemic variation in performance
    - Enhance front-desk staff and radiologist education and training with strategies, such as: (i) reviewing known cancer cases to enhance diagnostic pattern recognition, (ii) ongoing training on implicit bias and equity-focused modules, (iii) training staff to bring prior imaging for comparison, and (iv) promotion of double reading to improve diagnostic accuracy, which requires addressing the radiology workforce shortage
    - Integrate patient risk factors into recall decision-making (e.g., exploring the impact of dense breast tissue on recall rates and potential educational strategies around it)
    - Share clinician-specific data with providers and peers to promote performance awareness

- Conduct regular performance audits and create transparent reporting venues
- Complement recall rate with measures of timeliness to diagnostic follow-up, benign biopsy rates, and cancer detection to better contextualize recall quality and clinical outcomes
- Consider incentive structures (e.g., rewards for maintaining recall rates within a range) to promote sustained performance improvements (e.g., through pilot programs)

2.2.4 Next Steps

In the last session, Acumen provided a wrap up of the discussion and an overview of the next steps. After the meeting, Acumen distributed the OP-39 Re-specification TEP Poll to gather input from members on the discussions held during the webinar. Acumen will review the results and operationalize input for further testing of measure specification refinements based on the TEP meeting discussion and poll results and will follow up with TEP members with more information about the next steps.

Please contact **Acumen Outpatient Measures TEP Support** at op-measures-tep-support@acumenllc.com if you have any questions.