Partnership for Quality Measurement
Powered by Battelle

2026 Pre-Rulemaking Measure Review (PRMR) and Measure Set Review (MSR) Strategic Discussion & Learning Workshop

Dr. Michelle Schreiber | Centers for Medicare & Medicaid Services (CMS)

Brenna Rabel | Battelle

Jeff Geppert | Battelle

Dr. Meridith Eastman | Battelle

Dr. Lydia Stewart-Artz | Battelle

Kate Buchanan | Battelle

June 23-24, 2026

The analyses upon which this publication is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services. . Restricted: Use, duplication, or disclosure is subject to the restrictions as stated in Contract Number 75FCMC23C0010 between the Government and Battelle.

Partnership for Quality Measurement
Powered by Battelle

Day 1: Using Quality Measurement to Promote Wellness and Prevention, Patient Safety, and Health System-Level Readiness

June 23, 2026

Virtual Participation

We are pleased to welcome you and want to create a meaningful exchange.

To participate in the discourse, type questions or comments in the chat or raise your hand.

Battelle staff will serve as virtual moderators. Please unmute yourself when the moderators call on you.

Virtual Participation *(cont.)*

Please lower your hand and mute yourself following your question/comment.

If you are a call-in user, please state your first and last name before speaking.

If you experience technical issues, contact the project team via chat on the virtual platform or at PQMsupport@battelle.org.

Community Guidance

- Respect all voices and allow others to contribute.
- Remain engaged and actively participate.
- Keep your comments concise and focused.
- Share your experiences.
- Learn from others.

Partnership for Quality Measurement
Powered by Battelle

Welcome and Introductions

Brenna Rabel | Battelle

Meeting Agenda (Day 1)

1:00 PM	Welcome and Introductions
1:15 PM	CMS Opening Remarks
1:20 PM	Opening Session: Resetting the Frame
2:00 PM	Priority Problems Breakout Session
2:45 PM	Barriers and Data Gaps Breakout Session
3:15 PM	Design Session: What Could Quality Measurement Promote?
3:50 PM	Day 1 Reflections
4:00 PM	Adjourn

* All times listed in ET

7

Acronyms

- CBE: Consensus-Based Entity
- CMS: Centers for Medicare & Medicaid Services
- CMIT: CMS Measures Inventory Tool
- MSR: Measure Set Review
- FHIR®: Fast Healthcare Interoperability Resources®
- PA: Preliminary Assessment
- PIA: Performance and Impact Analysis
- PRMR: Pre-Rulemaking Measure Review
- PIE: Pre-Meeting Initial Evaluation
- PQM: Partnership for Quality Measurement
- USCDI: United States Core Data for Interoperability

Quality Measurement and Quality Improvement

Goals

Develop specific recommendations on how to better leverage our existing measures and data sources to address persistent quality problems

Consider how PRMR might best support the identification, evaluation, and recommendation of additional measures or data approaches that may be needed

Create a strategy for measurement that supports **accountability** and **improvement**

Partnership for Quality Measurement
Powered by Battelle

CMS Opening Remarks

Dr. Michelle Schreiber | CMS

Partnership for Quality Measurement
Powered by Battelle

Opening Session: Resetting the Frame

Brenna Rabel | Battelle

Day 1 Objective

Focus critical thinking about quality measurement to promote:

- Prevention and Wellness
- Patient Safety
- Health System Readiness

Patient Health Care Journey: Prevention Stages

Primary Prevention (Primary)

Prevents diseases and injuries before they occur. This objective is achieved by reducing unhealthy or unsafe behaviors that contribute to disease or injury and by increasing resistance to disease when exposure occurs.

Management of Acute Events, Chronic Disease, Functional Status (Tertiary)

Focuses on improving quality of life for individuals with acute and chronic conditions by minimizing or preventing the disease effects through quality care.

Initial Recognition & Management (Secondary)

Emphasizes timely recognition and diagnosis of conditions to enable effective treatments that work best in the early stages of illness and to support improved disease management overall.

Post-Acute Care, End-of-Life, Advanced Illness (Quaternary)

Post-acute care provides medical or supportive care to individuals after they leave an acute care setting, such as a hospital, but are not yet ready to return home. This care includes rehabilitation, home health care, and palliative care services.

Advanced illness care provides medical care for people living with serious illness. It addresses symptoms and illness-related stress and improves the overall experience and quality of life for patients, caregivers, and family members.

Patient Health Care Journey – Prevention Stages & Quality Problems

Patient Health Care Journey (Prevention Stage)	Addressable Quality Problems				
	Risk not visible early	Not actionable	Too late	Capacity vs effort unclear	Patient experience missing
Primary Prevention (Example)	**Data:** Claims + patient-reported (PR) **Tech:** Registries, FHIR exchange **Example:** Identifying rising cardiometabolic risk	**Data:** Clinical **Tech:** Clinical Decision Support, Electronic health record (EHR) prompts **Example:** Missed vaccination opportunities	**For discussion:** Can you think of how this quality problem might play out in primary prevention? Where might information and intervention be coming too late?	**Data:** Claims **Tech:** Not applicable **Example:** Low screening due to access	**Data:** PR **Tech:** Digital intake **Example:** Patient experience of care

Break

Please return by 2:00 PM.

Partnership for Quality Measurement
Powered by Battelle

Priority Problems Breakout Session

Based on your responses in the pre-meeting survey, you will be assigned to one of the following breakout groups:

- *Group 1: Delays in diagnosis or treatment*
- *Group 2: Fragmented care across settings*
- *Group 3: Preventable harm*
- *Group 4: Unwarranted variation in care delivery*

Problems in the Health Care System

What is the most important problem facing the health care system today in relation to wellness and prevention, health system readiness, or patient safety?

At which point in the patient health care journey does this problem occur?

Which setting or population is most affected?

Priority Problems Breakout Group 1: Delays in diagnosis or treatment

Jeff Geppert | Battelle

Breakout Session 1 Group 1

Discussion Questions:

At which point in the patient health care journey does this problem occur?

Which setting or population is most affected?

Priority Problem:

Delays in diagnosis or treatment

Priority Problems Breakout Group 2: Fragmented care across settings

Brenna Rabel | Battelle

Breakout Session 1 Group 2

Discussion Questions:

At which point in the patient health care journey does this problem occur?

Which setting or population is most affected?

Priority Problem:

Fragmented care across settings

Priority Problems Breakout Group 3: Preventable harm

Dr. Meridith Eastman | Battelle

Breakout Session 1 Group 3

Discussion Questions:

At which point in the patient health care journey does this problem occur?

Which setting or population is most affected?

Priority Problem:

Preventable harm

Priority Problems Breakout Group 4: Unwarranted variation in care delivery

Kate Buchanan | Battelle

Breakout Session 1 Group 4

Discussion Questions:

At which point in the patient health care journey does this problem occur?

Which setting or population is most affected?

Priority Problem:

Unwarranted variation in care delivery

Report Out:
Problem Themes

Break

Please return by 2:45 PM.

Partnership for Quality Measurement
Powered by Battelle

Barriers and Data Gaps Breakout Session

Based on your responses in the pre-meeting survey, you will be assigned to one of the following breakout groups:

- *Group 1: Delays in diagnosis or treatment*
- *Group 2: Fragmented care across settings*
- *Group 3: Preventable harm*
- *Group 4: Unwarranted variation in care delivery*

Barriers and Data Gaps

What is the main barrier preventing progress on this problem?

Which data sources currently exist but are underused or misused?

Which data sources are missing or arrive too late?

Barriers and Data Gaps Breakout Group 1: Delays in diagnosis or treatment

Jeff Geppert | Battelle

Breakout Session 2 Group 1

Discussion Questions:

- What is the main barrier preventing progress on this problem?
- Which data sources currently exist but are underused or misused?
- Which data sources are missing or arrive too late?

Priority Problem:

Delays in diagnosis or treatment

Barriers and Data Gaps Breakout Group 2: Fragmented care across settings

Brenna Rabel | Battelle

Breakout Session 2 Group 2

Discussion Questions:

- What is the main barrier preventing progress on this problem?
- Which data sources currently exist but are underused or misused?
- Which data sources are missing or arrive too late?

Priority Problem:

Fragmented care across settings

Barriers and Data Gaps Breakout Group 3: Preventable harm

Dr. Meridith Eastman | Battelle

Breakout Session 2 Group 3

Discussion Questions:

- What is the main barrier preventing progress on this problem?
- Which data sources currently exist but are underused or misused?
- Which data sources are missing or arrive too late?

Priority Problem:

Preventable harm

Barriers and Data Gaps
Breakout Group 4: Unwarranted variation in care delivery

Kate Buchanan | Battelle

Breakout Session 4 Group 4

Discussion Questions:

- What is the main barrier preventing progress on this problem?
- Which data sources currently exist but are underused or misused?
- Which data sources are missing or arrive too late?

Priority Problem:

Unwarranted variation in care delivery

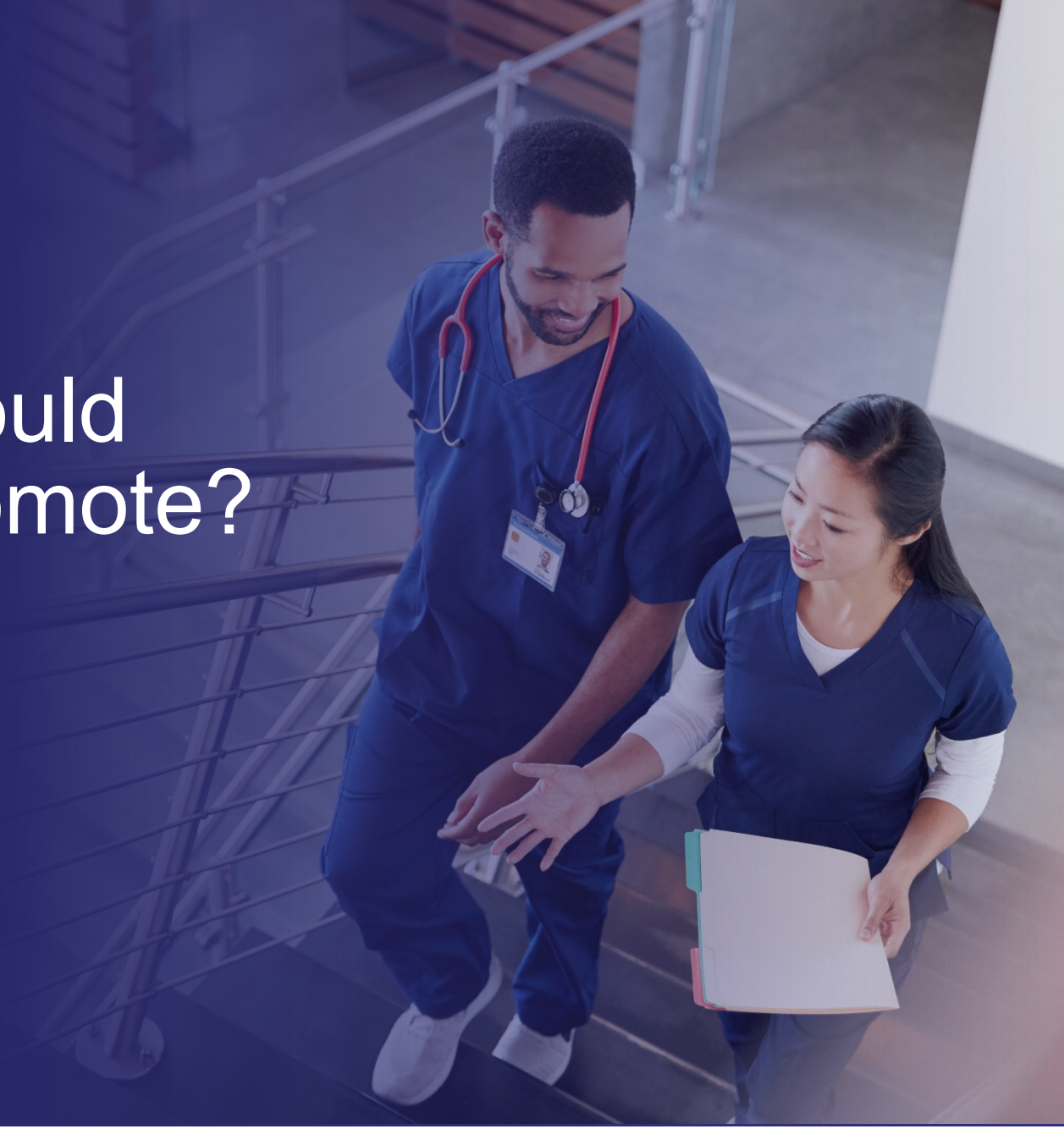
Report Out:
Barriers and Data Types

Partnership for Quality Measurement
Powered by Battelle

Design Session: What Could Quality Measurement Promote?

Dr. Meridith Eastman | Battelle

Jeff Geppert | Battelle



Early Intervention

- Which data types would support earlier intervention?
  - Which data types support early warning vs. retrospective assessment?
  - Which data types best support targeted assistance vs. accountability?
  - Where do combinations of data add most value for early intervention?

Consider data characteristics that support early intervention:
Data collection feasibility
Precision
Granularity
Timeliness
Clinical actionability
Sensitivity to change

Measure Examples

- What measures could support targeted help or readiness assessment?
  - Examples include measures that:
    - Identify areas where readiness is low, not only where performance is poor;
    - Provide data that support early warning signals;
    - Distinguish between inability to perform (“can’t do”) and failure to perform (“didn’t do”);
    - Use stratified data to target technical assistance and allocate resources;
    - Support learning across organizations;
    - Create feedback loops that make improvement visible.

Addressing Barriers Using Data and Measures

- How can quality data and measures help address the barriers discussed today?
- Where/how will measures NOT be helpful in addressing barriers?

Staffing constraints

Poor visibility into patient experience or functioning

Data lags that make data-informed quality improvement difficult

Limited ability to act on health-related risk before harm occurs

Inadequate information technology (including EHRs and interoperability)

Lack of timely, usable data to inform patient care

Misalignment between payment, reporting, and improvement efforts

Difficulty in distinguishing true differences in provider/facility performance from random differences

Partnership for Quality Measurement
Powered by Battelle

Day 1 Reflections

Brenna Rabel | Battelle

Partnership for
Quality Measurement
Powered by **Battelle**

Partnership for Quality Measurement
Powered by Battelle

Day 2: Learning Together—PRMR/MSR Evaluation Criteria Application & Resources to Support Measure Reviews

June 24, 2026

Virtual Participation

We are pleased to welcome you and want to create a meaningful exchange.

To participate in the discourse, type questions or comments in the chat or raise your hand.

Battelle staff will serve as virtual moderators. Please unmute yourself when the moderators call on you.

Virtual Participation *(cont.)*

Please lower your hand and mute yourself following your question/comment.

Please state your first and last name if you are a call-in user.

If you are experiencing technical issues, contact the project team via chat on the virtual platform or at PQMsupport@battelle.org.

Community Guidance

- Respect all voices and allow others to contribute.
- Remain engaged and actively participate.
- Keep your comments concise and focused.
- Share your experiences.
- Learn from others.

Partnership for Quality Measurement
Powered by Battelle

Welcome and Day 1 Recap

Dr. Meridith Eastman | Battelle

Meeting Agenda (Day 2)

1:00 PM	Welcome and Day 1 Recap
1:15 PM	Overview of PRMR and MSR Evaluation Criteria
1:45 PM	Scientific Acceptability Deep Dive
2:30 PM	Meaningfulness Applied Breakout Session
3:05 PM	PRMR/MSR Updates
3:25 PM	PRMR/MSR Resources for Committee Members
3:50 PM	Day 2 Reflections
4:00 PM	Adjourn

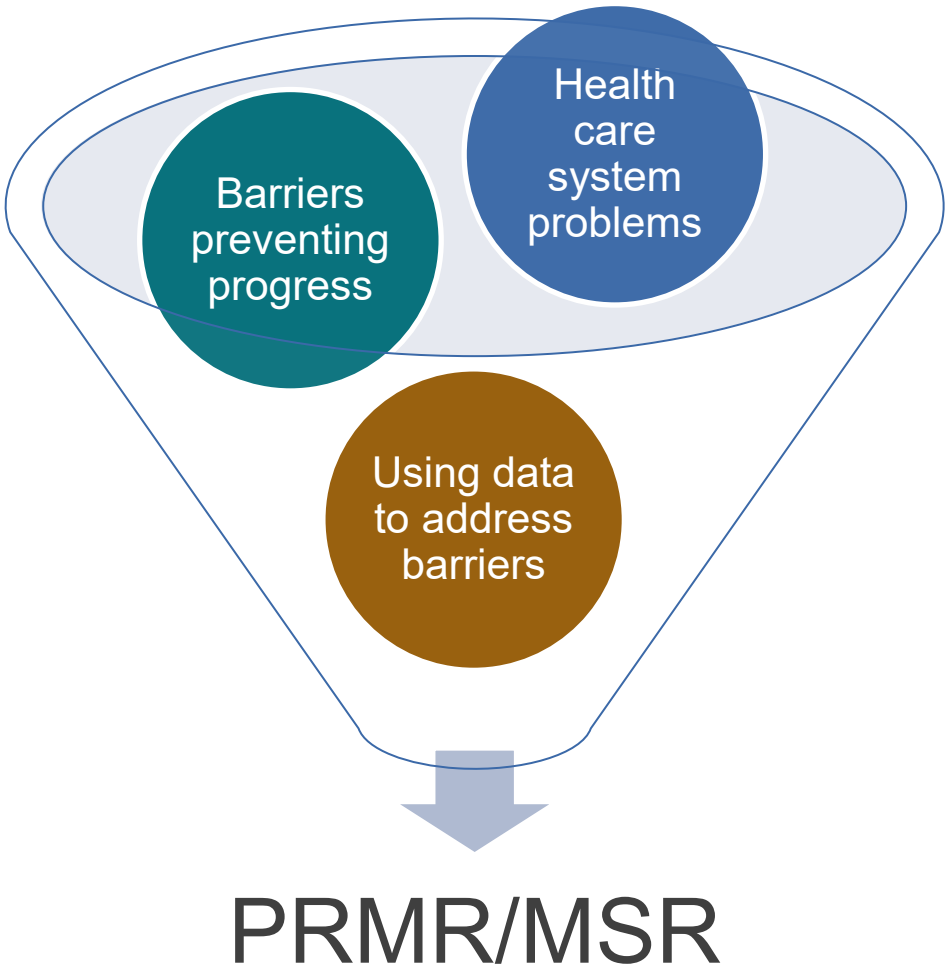
* All times listed in ET

50

Acronyms

- CBE: Consensus-Based Entity
- CMS: Centers for Medicare & Medicaid Services
- CMIT: CMS Measures Inventory Tool
- MSR: Measure Set Review
- FHIR®: Fast Healthcare Interoperability Resources®
- PA: Preliminary Assessment
- PIA: Performance and Impact Analysis
- PRMR: Pre-Rulemaking Measure Review
- PIE: Pre-Meeting Initial Evaluation
- PQM: Partnership for Quality Measurement
- USCDI: United States Core Data for Interoperability

Day 2 Objective



Today we will:

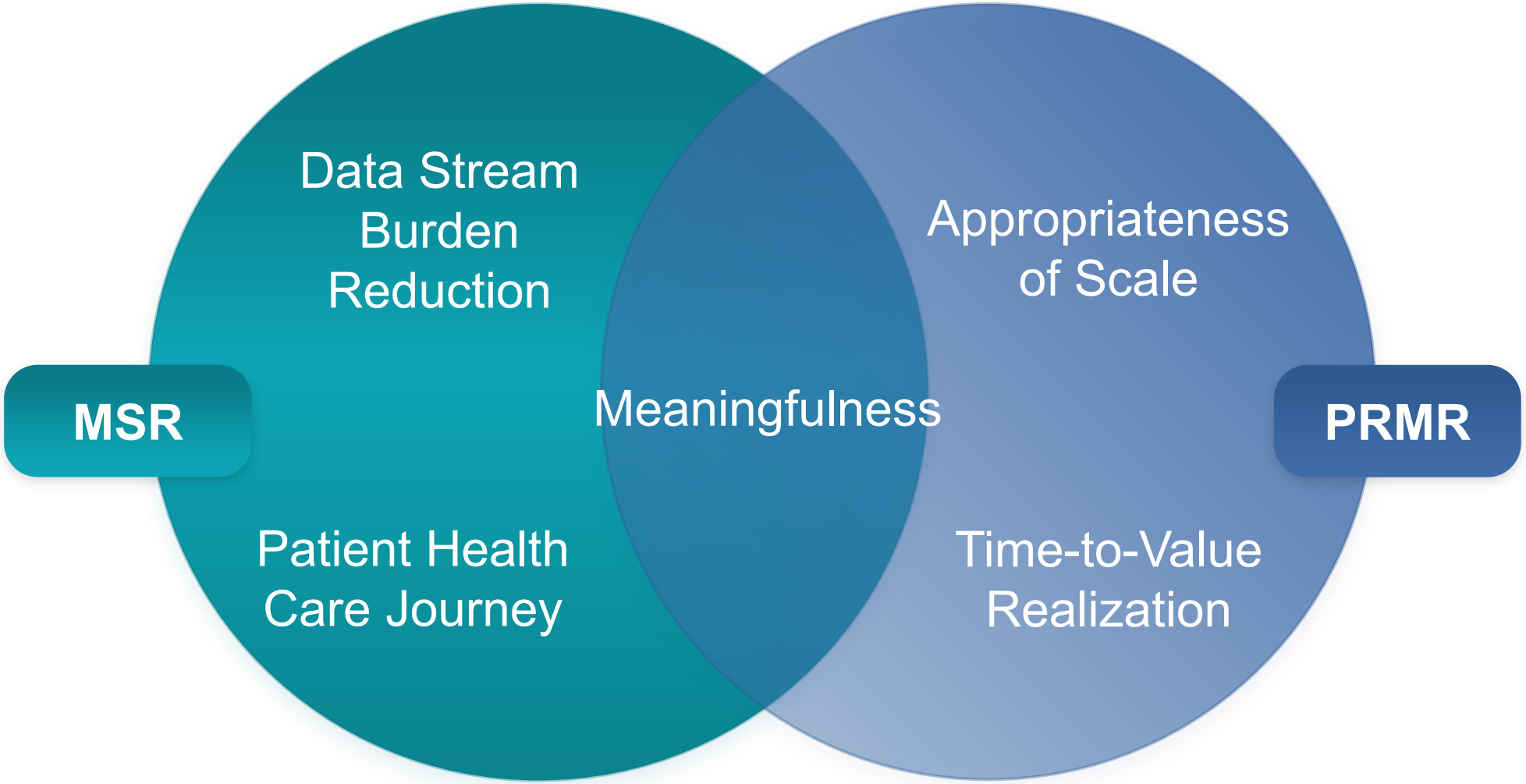
- Review PRMR/MSR evaluation criteria.
- Take a deep dive into scientific acceptability.
- Break out into groups to a review a sample PA.
- Review upcoming guidebook policy changes and schedules for PRMR and MSR.
- Tour the website and committee workspace.

Partnership for Quality Measurement
Powered by Battelle

Overview of PRMR and MSR Evaluation Criteria

Dr. Lydia Stewart-Artz | Battelle

PRMR & MSR Criteria



Meaningfulness – PRMR & MSR

Meaningfulness

Importance

Does this measure focus on an outcome that matters or on known gaps in care?

Feasibility

Are the tools, processes, and people necessary to implement and report on the measure reasonably available?

Conformance

Does the measure continue to function as intended when applied to current data sources, program populations, and settings?

Meaningfulness – PRMR & MSR *(cont.)*

Validity

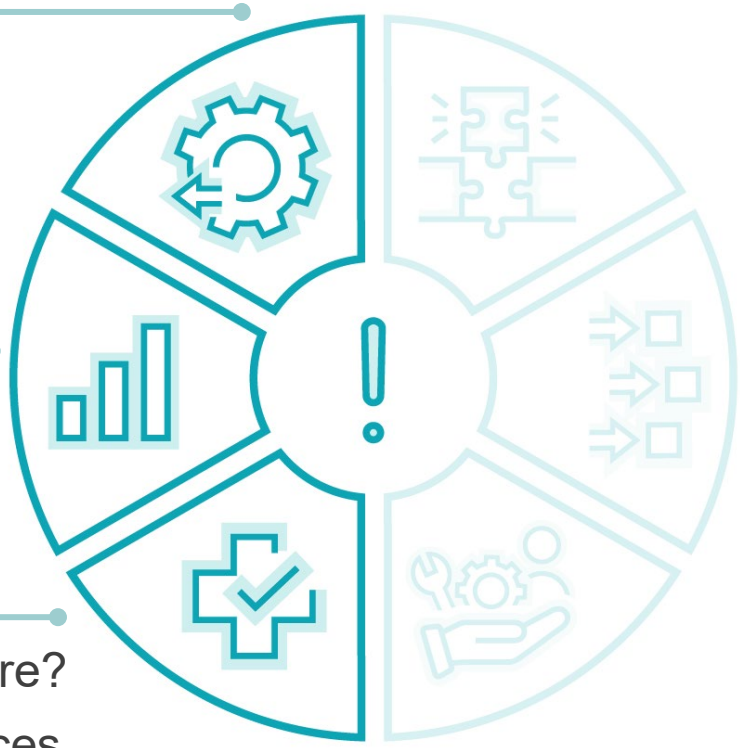
- Does the measure score accurately reflect quality of care?
- Are there clear and effective ways for reporting entities to improve performance on the measure?

Reliability

- Do differences in performance reflect true differences among reporting entities rather than random variation or measurement error?

Usability

- Does the measure provide actionable insights for improving quality of care?
- Have any potential barriers to implementation or unintended consequences been identified and mitigated?



Meaningfulness

Data Stream Burden Reduction – MSR

Data Stream Burden Reduction

When evaluating data stream burden reduction, committee members should evaluate whether:

- The clinical data flow required for the measure promotes **non-burdensome** data collection and reporting.

  Example: Does the measure require additional data workflows, separate submission mechanisms, or unique extraction processes that could be streamlined or eliminated?

- **Measure set redundancy** in data streams is identified and mitigated.

  Example: Is the measure duplicative of others in the set that assess similar gaps in care?

Patient Health Care Journey Alignment – MSR

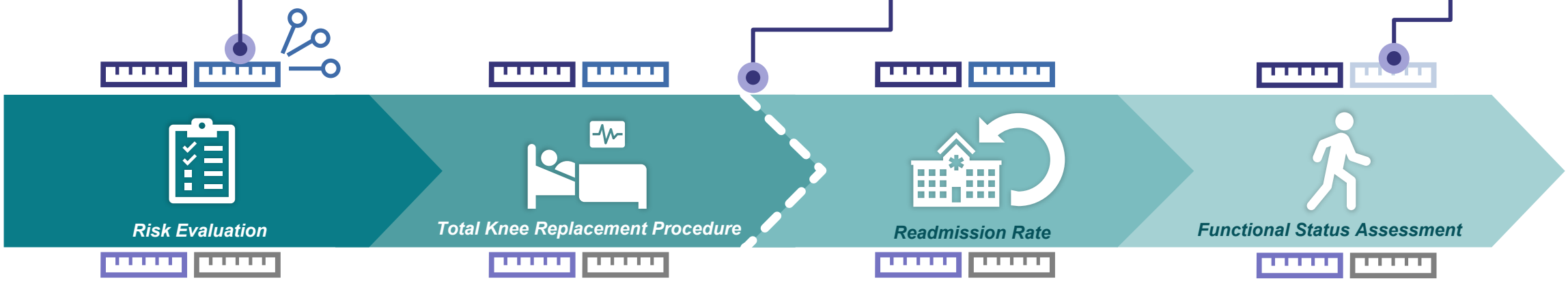
Alignment with the Patient Health Care Journey

When evaluating the patient journey, committee members should evaluate whether the measure aligns with how patients experience and prioritize care.

Does the measure address the right aspects of care for where patients typically are in their care journey?

Does the measure still help identify where gaps in care may occur along the care pathway?

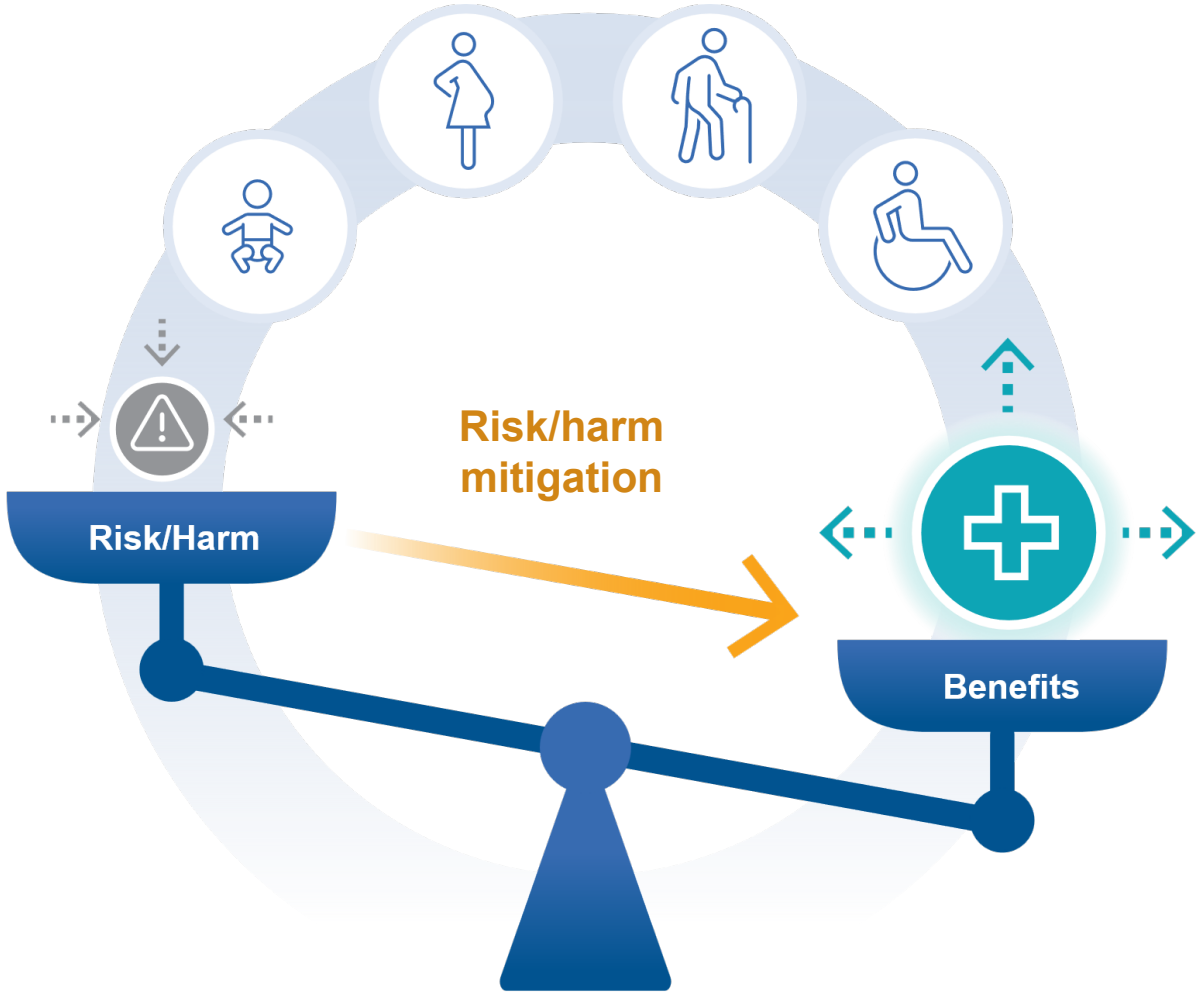
Would removing the measure leave an important part of the patient journey unmeasured, or does another measure now better cover that point in care?



Appropriateness of Scale – PRMR

Appropriateness of Scale:

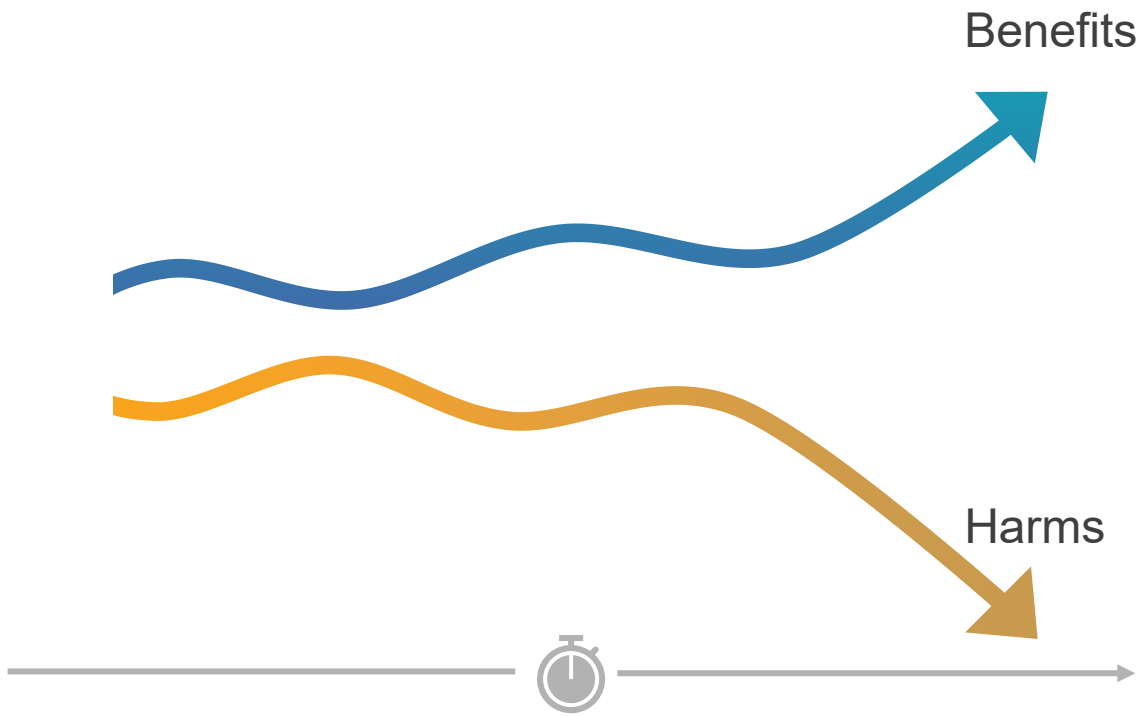
- Is the measure balanced and scaled to meet program-target population-specific goals?
  - ✓Evaluation of the appropriateness of scale assertion considers the evidence about the distribution of benefits and of risks/harms of the measure distributed across subpopulations and how risks/harms of the measure may be mitigated.



Time-to-Value Realization – PRMR

Time-to-Value Realization:

- Does the measure have a plan for near- and long-term positive impacts on the targeted program and population as the measure matures?
  - ✓Committees evaluate the time-to-value realization by considering how the harms and benefits change over time, ways the benefits of the measure might be prolonged, and how potential harms could be prevented.



Partnership for Quality Measurement
Powered by Battelle

Scientific Acceptability Deep Dive

Dr. Lydia Stewart-Artz | Battelle

Laura Aume | Battelle

Meaningfulness – Scientific Acceptability

Validity

- Does the measure score accurately reflect quality of care?
- Are there clear and effective ways for reporting entities to improve performance on the measure?

Reliability

- Do differences in performance reflect true differences among reporting entities rather than random variation or measurement error?

Meaningfulness

Introduction to Validity

What does validity mean for quality measures?

- A clinical quality measure is valid when it accurately captures the aspect of care it is designed to assess. In other words, the measure truly reflects the quality of care—not something else.

Why does it matter?

- The purpose of quality measures is to compare providers, inform patient choices, and determine payments.
- If a measure is not valid, those decisions could be based on misleading information—which may harm both patients and providers.
- Establishing a high degree of validity helps ensure that the measure is fair, meaningful, and useful in real-world settings.

A simple way to think about it:

- *Reliability asks: "Does this measure give consistent results?"*
- *Validity asks: "Does this measure capture what it claims to?"*

Why is Measuring Validity Important?

A measure with a strong degree of validity is one that truly captures what it claims to measure. The stronger the validity:

- The more confident we can be that people who pick a top-rated provider will actually receive better care—and for high-priority measures, see real improvements in their health.
- The more confident we can be that providers who work to improve measure scores will actually deliver higher-quality care.

- Keep in mind: what is true for a group on average may not hold for every individual patient—so we should be cautious about drawing broad conclusions.
- These best practices for establishing validity ensure measures are trustworthy enough to guide payment decisions and public quality reports.

Two Levels of Validity Testing

Person- or Encounter-Level

Key Questions: Does the measure specification accurately reflect the intent of the measure?

Focus: Unambiguous narrative measure specifications; data elements that accurately reflect the care process, procedure, or outcome measured; and value sets and codes that are complete, correct, and current.

Why It Matters: The measure specification, data elements, and value sets are the building blocks for a measure score. If they do not have a moderate to high degree of validity, measure score may not be accurate.

Common Activities and Tests :

- Delphi technique or nominal group method with a technical expert panel for narrative specification and value sets
- Gold standard testing
- Chance-adjusted agreement for data elements (Kappa, Gwet's AC1)
- Sensitivity and specificity of measure components
- Review of empirical literature

Accountable Entity-Level

Key Question: Do measure scores accurately reflect care processes, procedures, or outcomes received by members of measured groups?

Focus: A priori hypotheses about the degree or correlation with related and unrelated measures or processes, followed by testing. A causal mechanistic claim to explain the hypothesized relationship(s) is highly desirable.

Why It Matters: Hypothesis-driven testing, with supporting causal mechanistic claims, increase confidence that the measure score is an accurate reflection of quality.

Common Activities and Tests:

- A priori hypothesis generation
- Hypothesis-testing (convergent, discriminant validity evidence)
- Causal mechanistic claims aligned with logic model

Source: CMS MMS Hub — mmshub.cms.gov | Adams, 2009 (RAND TR-653)

Common Types of Validity Evidence

Face Validity **Subjective/Expert Judgment**

The extent to which a measure appears to measure what it is supposed to measure "at face value." Considered a weaker form because it is not based on objective data. Best practices use the Delphi technique or nominal group method with a technical expert panel (TEP).

Example: A TEP independently rates whether the measure specification for a blood pressure control measure will' accurately depict blood pressure control among patients. If not, the measure specification is revised with continual TEP review until consensus is reached (Delphi method).

Convergent Validity **Empirical/Data-Driven**

The measure score is highly correlated with scores from related measures.

Example: A diabetes care process measure (regular glucose testing at home) score if highly correlated with a diabetes outcome measure (HbA1c control), indicating that both capture aspects of diabetes care quality.

Discriminant Validity **Empirical/Data-Driven**

The measure score has a low correlation with scores from unrelated measures.

Example: A measure score reflecting child vaccination rates has a low correlation with scores from a measure of children with normal body mass index values, indicating that measure concepts are unique.

Criterion Validity **Empirical/Data-Driven**

The measure agrees with an established gold standard or benchmark for the same concept, providing strong evidence that it accurately captures what it intends to OR if at the data element level, data contributing to a measure score accurately reflect the care encounter.

Example: *An eCQM for blood pressure control captures electronically extracted blood pressure values that have been show to match (k=.82) those that are abstracted from the health record ('Gold standard").*

Source: CMS Measures Management System Hub — mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria/scientific-acceptability/validity

What is Reliability?

- **Reliability** is the degree to which a measure repeatedly and consistently produces the same result.[1]
- **Reliability & Stability:** Stability refers to the degree of variation over time (in either the performance score or the process the measure is intended to reflect).
  - It is possible (and often common) for a measure to be reliable but not stable.
  - When possible, developers are encouraged to track entity-level performance scores longitudinally to assess the measure's stability over time (i.e., the change in performance scores over time).
- **Reliability & Validity:** Reliability is a component of the validity argument. Specifically, it helps to rule out chance as a competing cause of variation, alongside other sources such as confounders and counteracting mechanisms.
  - A measure that is not reliable results in a validity judgement with considerable residual risk.

1. International Organization for Standardization. (2019). Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality measurement framework (ISO/IEC Standard No. 25020:2019).

Evaluation of Reliability

Clinical quality and cost/resource use measures are evaluated at two levels:

- Person/encounter level
  - Generally defined as the absence of ambiguity in the measure specification and the repeatability of the data collection process for elements such as diagnoses, procedures, or lab results recorded across different encounters and individuals.
  - High reliability at this level ensures that the underlying data used to construct the measure are consistent and reproducible.
- Accountable entity level
  - Assesses whether observed differences in performance reflect true quality rather than random variation.
  - High reliability at this level ensures that comparisons across entities are based on stable and repeatable results, not on random fluctuations or inconsistent data.

Two Levels of Reliability Testing

Person/Encounter Level

Data Element Reliability

Key Question: "Are the individual data points that make up a measure collected consistently and without ambiguity?"

Focus: Repeatability of data collection—diagnoses, procedures, lab values, survey responses—at the individual patient or encounter level.

Why It Matters: If the raw data going into a measure are unreliable, no amount of aggregation can fix the result. This is the foundation of the measure.

Common Measures & Example Tests :

- Test-Retest Reliability (ICC, Correlation Coefficient)
- Inter- and Intra-Rater Agreement (ICC, Kappa, Percent Agreement)
- Internal Consistency (Cronbach's alpha)

Example: *Two chart abstractors review the same 100 records to determine whether patients received recommended therapy. A Cohen's Kappa of 0.82 shows they reach the same conclusion reliably.*

Accountable Entity Level

Measure Score Reliability

Key Question: "Can we reliably distinguish between high-performing and low-performing hospitals, clinicians, or health plans?"

Focus: The ratio of "signal to noise" in computed measure scores (Adams, 2009). Signal = true quality differences between entities. Noise = random variation.

Why It Matters: Low entity-level reliability increases the probability of misclassifying a provider's performance

Common Measures and Example Tests:

- Signal-to-Noise (Beta-Binomial Model)
- Temporal Correlation (ICC, Pearson's correlation coefficient, Spearman's p, Kendall's Tau)
- Random Split-Half Correlation (ICC, Pearson's correlation coefficient, Spearman's p, Kendall's Tau)

Example: *A hospital readmission measure is computed for 300 hospitals. A signal-to-noise reliability of 0.85 at the median means observed differences mostly reflect true quality variation, not chance.*

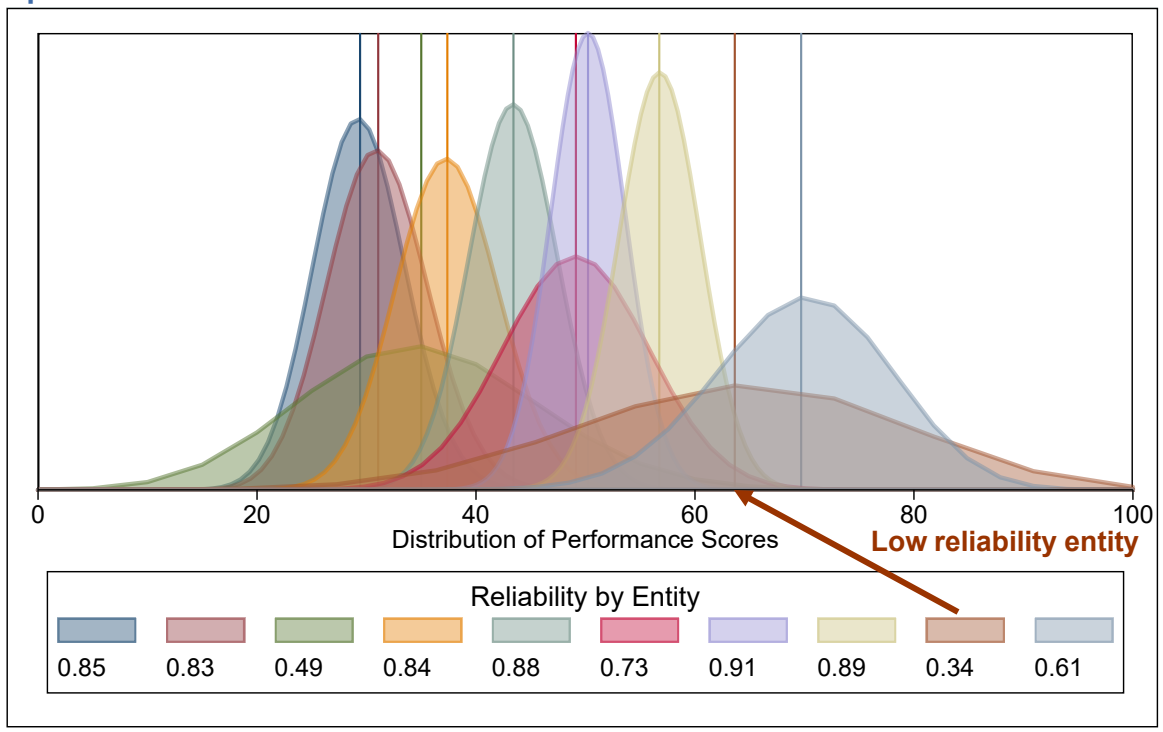
Key Takeaway: Person/encounter reliability ensures the building blocks are solid. Entity-level reliability ensures the final scores can fairly compare providers.

Source: CMS MMS Hub — mmshub.cms.gov | Adams, 2009 (RAND TR-653)

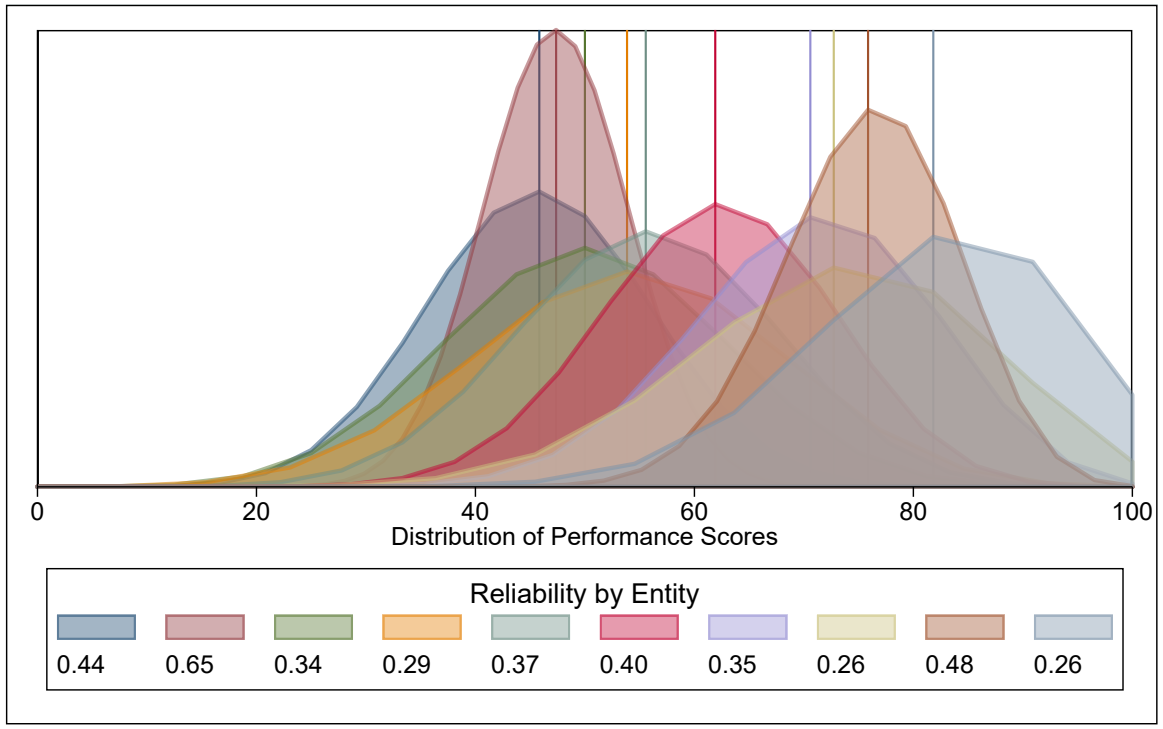
Entity-Level Reliability

- Entity-level reliability measures the ability to distinguish an entity's performance from other entities.

High overall reliability: can distinguish levels of performance between entities



Low overall reliability: overlap across most entities



Accountable Entity-Level Reliability Methods at a Glance

Signal-to-Noise Reliability

What we ask:

"How much of the difference between entities is real quality variation vs. random noise?"

Use when:

You have yes/no patient-level data (e.g., "Did the patient receive the recommended treatment?") and want to confirm that score differences between entities are meaningful, not just chance.

Think of it as:

The go-to method for most measures with binary (yes/no) data. It directly estimates how much of the observed variation is "signal" (true differences) vs. "noise" (random).

Key test: Beta-binomial model (Adams, 2009)

Random Split-Half Correlation

What we ask:

"If we split each entity's patients into two random groups, do we get the same score both times?"

Use when:

An independent dataset or data from an adjacent time period are not available. This is the fallback when temporal correlation is not possible.

Think of it as:

An internal consistency check. You randomly deal each entity's patients into two piles, score both piles, and see if the answers agree. If they do, the measure is stable enough to trust.

Key tests: ICC, Pearson's r, Spearman's ρ, Kendall's τ; Spearman-Brown Prophecy Formula

Temporal Correlation

What we ask:

"If we measure the same entities in two consecutive periods, do the rankings stay about the same?"

Use when:

You want to compare a dataset from the same source separated by a minimal time period (e.g., Q1 vs. Q2, or Year 1 vs. Year 2) to verify scores are stable over time.

Think of it as:

A "test-retest" at the entity level. If a hospital scored well last quarter, does it still score well this quarter? High correlation means entity performance is consistent, not a fluke.

Key tests: ICC, Pearson's r, Spearman's ρ, Kendall's τ; Spearman-Brown Prophecy Formula

Source: CMS MMS Hub — mmshub.cms.gov

Key Considerations for Validity and Reliability

Population Representativeness

- Testing population should match the intended program population (age, geography, payer mix).
- Results must be generalizable to the target population.
- Check for differential performance across subgroups (rural vs. urban, small vs. large entities).

Sample Size and Adequacy

- Sample sizes must support stable reliability estimates.
- Entity-level reliability is sensitive to denominator size—report results by decile when possible.
- Include enough entities to represent the full range of performance.

Testing Method Appropriateness

- Match the test to the data type (Kappa for categorical, ICC for continuous, beta-binomial for dichotomous).
- Test at both levels: person/encounter and accountable entity.
- Go beyond descriptive statistics—provide true reliability and validity evidence.

Risk Adjustment and Stratification

- Risk factors should be clinically appropriate and evidence based.
- Stratify to detect differential performance by race, ethnicity, or socioeconomic status.
- Ensure the measure does not penalize entities serving high-risk or underserved populations.

Data Source and Quality

- Data source (claims, electronic health record [EHR], registry, survey) must be appropriate and validated for the measure.
- How missing data are handled should not introduce systematic bias.
- Exclusions must be clinically justified and documented.

Re-specification and Maintenance

- Respecified measures (new population, setting, or data source) require retesting.
- Maintenance measures need current testing—clinical practice or coding changes may affect validity.
- Face validity alone is not sufficient for a fully developed measure.

Source: CMS MMS Hub — Scientific Acceptability Evaluation Criteria

Guiding Questions

- **Importance**
  - Is there evidence that the measure focus is associated with a measurable and significant outcome or known gaps in care for persons and entities?
- **Conformance**
  - Do measure components and specifications align with the intent of the measure focus and target population?
- **Feasibility**
  - Are the tools, processes, and people necessary to implement and report on the measure reasonably available?
- **Usability**
  - Once implemented, will the measure provide actionable insights for improving quality of care?
  - Have all potential barriers to implementation or unintended consequences of use been identified?

Validity

- **Validity**
  - Did the developer show with data or reasoning that the measure accurately reflects quality of care in the program's setting?
  - Consider whether the measure captures true quality differences, not differences driven by factors outside a provider's control.
    - **Rule-Out:** Did the developer credibly rule out alternative explanations for observed performance differences (e.g., patient mix, random variation, data artifacts)?
    - **Rule-In:** Did the developer credibly demonstrate specific ways in which provider actions can influence measured performance?
  - These considerations help assess the risk of misclassification.

Reliability

- **Reliability**
  - Did the developer provide entity-level (measure score) testing showing that performance differences reflect true differences instead of measurement error?
    - Low reliability increases the risk that providers/entities are misclassified as lower performers when their true performance is average.
    - Reliability is considered acceptable when at least 70% of entities demonstrate reliability greater than 0.6.

Breakout Exercise: Sample Measure Case Studies

- Each breakout group will review a **fictional** quality measure and evaluate it against the meaningfulness criteria. Use the discussion prompts from the preceding slides to guide your assessment.
- Task: Decide if your measure would receive a “met” or “not met” on each of the sub-criteria within meaningfulness.

- **Group 1:** Timely Sepsis Bundle Compliance Rate (Patient Safety)
- **Group 2:** Postpartum Depression Screening and Follow-Up (Prevention & Wellness)
- **Group 3:** Chronic Condition Care Coordination Composite (Fragmented Care)
- **Group 4:** Hospital Antibiotic Stewardship Composite (Health System Readiness)

Break

Please return by 2:30 PM.

Partnership for Quality Measurement
Powered by Battelle

Meaningfulness Applied Breakout Session

Please wait to be assigned to one of the following breakout rooms:

- *Group 1: Timely Sepsis Bundle Compliance Rate (Patient Safety)*
- *Group 2: Postpartum Depression Screening and Follow-Up (Prevention & Wellness)*
- *Group 3: Chronic Condition Care Coordination Composite (Fragmented Care)*
- *Group 4: Hospital Antibiotic Stewardship Composite (Health System Readiness)*

Meaningfulness Applied Breakout Group 1: Timely Sepsis Bundle Compliance Rate (Patient Safety)

Dr. Lydia Stewart-Artz | Battelle

Group 1: Timely Sepsis Bundle Compliance Rate

- *This process measure assesses the percentage of adult patients (age 18+) presenting to the emergency department with suspected sepsis who receive all elements of an evidence-based treatment bundle within 3 hours of clinical recognition. The bundle includes: (1) blood cultures drawn prior to antibiotic administration, (2) serum lactate measurement, (3) administration of broad-spectrum antibiotics, and (4) initiation of IV fluid resuscitation for patients with hypotension or lactate ≥ 4 mmol/L. The measure uses EHR-extracted clinical data and is proposed for the Hospital Inpatient Quality Reporting (IQR) Program.*
- **Importance**
  - Sepsis is the leading cause of in-hospital mortality in the U.S., accounting for over 350,000 deaths annually and $62 billion in health care costs. Evidence from the Surviving Sepsis Campaign demonstrates that each hour of delayed antibiotic administration increases mortality risk by 4-8%.
  - A documented performance gap exists: national data show only 55-65% of sepsis patients receive all bundle elements within the recommended timeframe. The gap is most pronounced in smaller community hospitals and rural critical access hospitals, where sepsis recognition may be delayed due to lower case volumes and limited specialist availability.
- **Conformance**
  - The measure specifications align with the 2021 Surviving Sepsis Campaign international guidelines and target the adult inpatient population admitted through the ED. The four bundle components are well-established, evidence-based interventions.
  - A potential concern: the 3-hour compliance window is applied uniformly regardless of clinical presentation complexity. Some clinicians argue that patients with ambiguous early presentations may require longer diagnostic workups before sepsis is confirmed, potentially penalizing thoughtful clinical decision-making.
  - The measure excludes patients with comfort care orders and those transferred from other acute care facilities, which aligns with the intent to capture initial recognition and treatment.

Group 1: Timely Sepsis Bundle Compliance Rate *(cont., 1)*

- **Feasibility**
  - The developer submitted eCQM feasibility testing demonstrating successful data extraction from three major EHR platforms (Epic, Cerner, MEDITECH) across 12 pilot sites. Most acute care hospitals already capture the required data elements in structured fields.
  - However, the developer identified feasibility challenges at rural and critical access hospitals: 23% of pilot sites reported that point-of-care lactate testing was not consistently available in the ED, and real-time lactate results were sometimes recorded in free-text notes rather than discrete fields, reducing extraction accuracy.
- **Validity**
  - The developer provided convergent validity evidence showing correlation between bundle compliance rates and reduced in-hospital sepsis mortality ($r = 0.67$, $p < 0.001$) across 400 hospitals. A 12-member TEP established face validity using a modified Delphi process with two rounds of independent rating.
  - The developer conducted risk adjustment for age, comorbidity burden (Elixhauser index), and initial illness severity (SOI). However, the risk model does not include patient transfer status and time-of-day presentation (night/weekend vs. daytime). These unaddressed factors could drive observed performance differences.
  - The testing population was drawn from urban academic and community hospitals. Rural and critical access hospitals were underrepresented (8% of sample vs. 28% nationally).

Group 1: Timely Sepsis Bundle Compliance Rate *(cont., 2)*

- **Reliability**
  - Signal-to-noise analysis using a beta-binomial model across 400 hospitals showed reliability of 0.72 at the median. Approximately 74% of entities demonstrated reliability greater than 0.6, meeting the 70% acceptability threshold.
  - However, when stratified by hospital volume, facilities with fewer than 25 eligible sepsis cases per year showed substantially lower reliability (median 0.41), raising concerns about misclassification for low-volume hospitals.
  - The developer provided a misclassification analysis showing that 18% of low-volume hospitals would be incorrectly classified as below-average performers. Temporal correlation across two adjacent measurement periods showed a Pearson r of 0.81, supporting score stability over time.
- **Usability**
  - The measure provides clear, actionable improvement pathways: hospitals can implement sepsis screening protocols, create rapid response teams, conduct staff education, and deploy EHR-based clinical decision support alerts.
  - Potential unintended consequence: providers may feel pressured to administer broad-spectrum antibiotics to patients who meet screening criteria but do not ultimately have confirmed infection, contributing to antibiotic overuse. The developer acknowledged this risk but did not provide data quantifying its frequency.

Meaningfulness Applied Breakout Group 2: Postpartum Depression Screening and Follow-Up (Prevention & Wellness)

Amanda Overholt | Battelle

Group 2: Postpartum Depression Screening & Follow-Up

- *This measure calculates the percentage of patients who delivered a live birth during the measurement period who were (1) screened for postpartum depression using a validated tool (PHQ-9 or EPDS) within 12 weeks of delivery and (2) if screened positive (score ≥ 10), received a documented follow-up plan within 30 days. The measure uses EHR-extracted data and is proposed for the Merit-based Incentive Payment System (MIPS). The developer did not submit eCQM feasibility testing.*
- **Importance**
  - Postpartum depression affects approximately one in seven new mothers and is the most common complication of pregnancy. It is associated with impaired maternal-infant bonding, infant developmental delays, increased risk of maternal self-harm, and long-term behavioral health consequences for children.
  - National screening rates remain below 80%, with significant gaps in care: only 51% of Medicaid-enrolled mothers receive documented screening compared to 74% of commercially insured mothers. Follow-up rates after a positive screen are estimated at only 40-55% nationally.
- **Conformance**
  - The measure aligns with American College of Obstetricians and Gynecologists (ACOG) Committee Opinion 757 and United States Preventive Services Taskforce (USPSTF) B-recommendation for universal perinatal depression screening. The two-component structure (screening plus follow-up) reflects the full clinical pathway.
  - Committee members should consider: 25-30% of postpartum depression cases onset between weeks 12 and 26, potentially falling outside the measure's capture window. The measure also does not specify minimum elements for a "documented follow-up plan," leaving room for variation across practices.
  - Additionally, the measure does not clarify whether screening during any encounter (e.g., pediatric well-child visit with co-located services) qualifies, or only at a dedicated postpartum visit.

Group 2: Postpartum Depression Screening & Follow-Up *(cont., 1)*

- **Feasibility**
  - Screening tools (PHQ-9, EPDS) are widely available and already integrated into many EHR systems. The screening component is generally feasible for most obstetric practices.
  - However, significant feasibility concerns exist for the follow-up component. It requires linking data across obstetric and behavioral health settings, which depends on interoperability infrastructure that is inconsistent across health systems. In fragmented care environments, follow-up documentation may reside in a behavioral health record not accessible to the reporting entity.
  - The developer did not submit eCQM feasibility testing. For eCQMs, feasibility testing is required per Measures Management System (MMS) Hub guidance. This is a notable gap in the submission.
- **Validity**
  - A 10-member TEP established face validity using a modified Delphi process. Criterion validity evidence compares EHR-extracted screening data against chart-abstracted records, showing substantial agreement for the screening component (kappa = 0.78).
  - However, criterion validity for the follow-up component is weaker (kappa = 0.59), largely because abstractors inconsistently interpreted “documented follow-up plan.”
  - The developer does not address whether observed performance differences might reflect patient engagement barriers (stigma, transportation, childcare), insurance-driven access limitations, or regional behavioral health workforce shortages rather than provider quality.

Group 2: Postpartum Depression Screening & Follow-Up *(cont., 2)*

- **Reliability**
  - Split-sample reliability testing across 250 clinician groups shows an ICC of 0.65. Only 62% of entities demonstrates reliability above 0.6, falling below the 70% acceptability threshold.
  - The developer attributes the low reliability to small denominator sizes in solo and small group practices, where annual delivery volumes may be as low as 15-30 patients. At this volume, a small number of missed screenings dramatically shift the performance rate.
  - The developer does not translate reliability results into misclassification risk estimates as recommended by MMS Hub guidance. Without this analysis, the committee cannot assess how many entities would be incorrectly classified as low performers.
- **Usability**
  - The measure supports improvement through workflow redesign (integrating screening into standard postpartum visit protocols), care coordination (warm handoffs to behavioral health), and system investment (co-locating behavioral health in obstetric settings).
  - A significant unintended consequence concern: providers may increase screening rates without corresponding investment in follow-up infrastructure, creating a “screen and abandon” pattern. Research suggests that identifying depression without providing adequate support increases patient distress and erodes trust in the health care system.
  - This concern is amplified in settings with existing behavioral health workforce shortages, where the measure may penalize practices that cannot access follow-up services regardless of their screening quality.

Meaningfulness Applied Breakout Group 3: Chronic Condition Care Coordination Composite (Fragmented Care)

Dr. Abigail Evans | Battelle

Group 3: Chronic Condition Care Coordination Composite

- *This composite measure evaluates care coordination for patients aged 65+ with three or more chronic conditions across inpatient, outpatient, and post-acute settings. It assesses: (1) documented shared care plan accessible to all treating providers within 48 hours of discharge, (2) follow-up visit with a primary or specialty provider within 7 days of hospital discharge, and (3) medication reconciliation at each care transition. The measure combines claims and EHR data and is proposed for the Hospital Value-Based Purchasing (VBP) Program.*
- **Importance**
  - Patients with multiple chronic conditions account for over 70% of Medicare spending. Care fragmentation across settings is a well-documented driver of preventable hospitalizations, adverse drug events, and duplicative testing.
  - Studies estimate that poor care coordination contributes to 20-30% of avoidable readmissions in this population. Medication errors during care transitions affect an estimated 66% of patients with three or more chronic conditions, with 20-25% classified as potentially harmful.
- **Conformance**
  - The three-component structure reflects the full scope of care coordination per the National Quality Forum (NQF) framework. Each component targets a distinct, evidence-based element of effective care transitions.
  - However, the composite combines two process elements (care plan, medication reconciliation) with a timeliness element (7-day follow-up), and these may have different sensitivity to provider action. The follow-up visit may depend more on patient factors and outpatient scheduling than on hospital discharge quality.
  - The specification does not clearly define “shared” care plan. Is a plan “shared” if documented in an EHR with portal access, or must it be actively transmitted via a care summary document?

Group 3: Chronic Condition Care Coordination Composite *(cont., 1)*

- **Feasibility**
  - The claims-based components (follow-up visit, medication reconciliation codes) are straightforward and use existing billing infrastructure.
  - The shared care plan component presents significant feasibility challenges. It depends on health information exchange (HIE) infrastructure that varies dramatically by region. A 2024 Office of the National Coordinator for Health Information Technology (ONC) report found that only 42% of hospitals participate in a functioning HIE network, with rates as low as 18% in rural states.
  - Safety-net hospitals and those serving Medicaid-heavy populations report the lowest interoperability capacity. The developer did not test extraction of the “shared care plan” element across diverse EHR platforms.
- **Validity**
  - The developer demonstrates convergent validity with 30-day all-cause readmission rates ($r = –0.54$, $p < 0.001$) across 350 hospitals.
  - Developers explore risk adjustment during testing through a risk model. The model adjusts for patient age, comorbidity burden, and dual-eligibility status but does account for geographic access to post-acute care, community-level social determinants (transportation, broadband for telehealth), or regional HIE maturity. These systematically disadvantage hospitals in under-resourced areas.

Group 3: Chronic Condition Care Coordination Composite *(cont., 2)*

- **Reliability**
  - Temporal correlation across 2 consecutive years for 350 hospitals yielded a Pearson r of 0.79. Signal-to-noise reliability was 0.74 at the median, with 76% of entities above 0.6, meeting the 70% acceptability threshold.
  - However, individual component reliability varied substantially. Medication reconciliation (0.81) and follow-up visit (0.77) showed strong reliability, while the shared care plan component showed notably lower reliability (0.58) due to inconsistent interpretation of “shared” across sites.
- **Usability**
  - The measure provides actionable improvement targets: implementing care transition programs, investing in HIE connectivity, deploying pharmacist-led medication reconciliation, and creating standardized discharge workflows.
  - A critical closing gaps in care concern: hospitals serving patients discharged to areas with limited post-acute resources (few home health agencies, long wait times for outpatient appointments) may be penalized for the 7-day follow-up component regardless of discharge quality. Safety-net and rural hospitals could be systematically disadvantaged.
  - The developer did not identify or address this barrier in the submission, nor propose mitigation strategies such as exclusions for geographic access limitations or stratified reporting.

Meaningfulness Applied Breakout Group 4: Hospital Antibiotic Stewardship Composite (Health System Readiness)

Laura Aume | Battelle

Group 4: Hospital Antibiotic Stewardship Composite

- *This composite measure evaluates hospital-level antibiotic stewardship across three components: (1) percentage of surgical patients receiving guideline-concordant prophylactic antibiotic selection and timing, (2) percentage of antibiotic orders reviewed by a pharmacist or stewardship team within 48 hours, and (3) facility-level standardized infection ratio (SIR) for hospital-onset C. diff infection. Components one and two use EHR data; component three uses National Healthcare Safety Network (NHSN) surveillance data. Proposed for the Hospital Inpatient Quality Reporting Program (Hospital IQR).*
- **Importance**
  - The Centers for Disease Control and Prevention (CDC) has identified antibiotic resistance as one of the most urgent public health threats, causing over 2.8 million resistant infections and 35,000 deaths annually. Approximately 30% of inpatient antibiotic prescriptions are unnecessary or suboptimal.
  - Hospital-onset C. diff infections cause an estimated 12,800 deaths annually, with inappropriate antibiotic use as the primary modifiable risk factor. Despite Joint Commission stewardship requirements, no existing CMS measure comprehensively assesses stewardship practice.
- **Conformance**
  - The composite structure links process measures (prophylactic selection, stewardship review) with an outcome measure (C. diff SIR) to capture the full stewardship pathway.
  - However, C. diff rates are influenced by factors well beyond antibiotic prescribing: infection prevention practices (hand hygiene, environmental cleaning), patient comorbidities (proton pump inhibitor (PPI) use, immunosuppression), community transmission patterns, and regional strain virulence. Combining a multi-causal outcome with targeted process measures in a single composite may obscure the relationship between components.
  - The 48-hour pharmacist review window was based on expert consensus only. The developer acknowledged that no studies validate 48 hours as a clinically meaningful threshold versus 24 or 72 hours.

Group 4: Hospital Antibiotic Stewardship Composite *(cont., 1)*

- **Feasibility**
  - Component 3 (C. diff SIR) uses existing NHSN reporting infrastructure and is highly feasible. Component 1 (prophylactic antibiotic selection) can be extracted from structured EHR medication and procedure records at most hospitals.
  - Component 2 (48-hour pharmacist review) presents significant barriers. It requires structured documentation of stewardship review activities not standardized across EHR platforms. Of hospitals under 200 beds, 35% report having no formal stewardship program beyond minimum Joint Commission requirements.
  - The developer submitted feasibility testing from eight pilot sites (all large academic medical centers with established stewardship programs). Testing did not include community or rural hospitals.
- **Validity**
  - Convergent validity evidence showed higher composite scores correlate with lower overall antibiotic utilization rates ($r = -0.61$). Rule-in testing demonstrated that hospitals with established stewardship programs scored 22% higher on average.
  - The developer risk-adjusted Component 3 using the NHSN standardization model for facility case mix. However, community-level antibiotic resistance patterns, regional C. diff prevalence, and facility infection control investment (beyond stewardship) were not accounted for.
  - For Components 1 and 2, no risk adjustment was applied. The developer argued these are “all-or-nothing” process measures not affected by patient factors. This reasoning may be valid for prophylactic selection but is debatable for the pharmacist review component, which depends on institutional staffing and infrastructure.

Group 4: Hospital Antibiotic Stewardship Composite *(cont., 2)*

- **Reliability**
  - Entity-level signal-to-noise reliability is 0.69 at the median across 500 hospitals, with 68% of entities above the 0.6 threshold. This falls below the 70% acceptability standard.
  - Component-level analysis reveals the source: the two process components individually show strong reliability (prophylactic selection: 0.82; pharmacist review: 0.76), but the C. diff SIR outcome component shows reliability of only 0.52 due to low event rates. Hospitals with fewer than 10 C. diff events per year (approximately 40% of the sample) have a median outcome reliability of 0.34.
  - The developer does not provide misclassification risk estimates that the MMS Hub guidance recommends. A single component pulls the composite's overall reliability below threshold.
- **Usability**
  - The measure supports concrete improvement actions: formulary optimization, stewardship program staffing, surgical prophylaxis protocol standardization, and pharmacist-led prospective audit and feedback. These have strong evidence bases in the infectious disease literature.
  - Potential unintended consequences: (1) providers may under-prescribe antibiotics in clinically ambiguous situations to improve process scores, potentially delaying necessary treatment; (2) the stewardship review component may incentivize “rubber stamp” reviews that meet documentation requirements without meaningful clinical assessment.
  - The developer acknowledges both risks but does not provide data quantifying their likelihood or propose mitigation strategies.

Report Out:
Meaningfulness Applied

Break

Please return by 3:05 PM.

Partnership for Quality Measurement
Powered by Battelle

PRMR/MSR Updates

Dr. Meridith Eastman | Battelle

Kate Buchanan | Battelle

PRMR/MSR Updates

1 PRMR Recommendations Meeting Structure

The meeting now includes both the Advisory Group and Recommendation Group. The Advisory Group provides their input first, and then the Recommendation Group discusses and votes.

2 PRMR Measure Previews

Battelle will host virtual previews to help committee members prepare for their PRMR Recommendations Meeting. Advisory and Recommendation Group members may ask clarifying questions about Measures Under Consideration (MUCs).

3 MSR Performance and Impact Analyses

Each of the up to 50 preliminary MSR measures will have a Performance and Impact Analysis that references publicly available performance data and other information about measure characteristics.

Updates

PRMR/MSR Updates *(cont., 1)*

4

PQM Website Workspace Enhancements

The updated workspace centralizes member details and features a discussion board, resources, news, and events.

5

Public Comment Posting Policy

Battelle will post public comments, as written, by the end of the public comment period unless otherwise noted. Commenters should submit only information they are comfortable making public and avoid including prohibited content.

Updates

MSR Schedule

August

- **Aug. 6:** MSR Patient Education Meeting
- **Aug. 5:** MSR Committee Member Education Meeting

September

- **Sep. 18:** PIE Forms open
- **Sep. 18:** Final PAs and list of MSR measures published

October

- **Oct. 2:** PIE Forms due back from committee members
- **Oct. 13-14:** MSR Recommendations Meeting

November

- **Nov. 25:** Final MSR Recommendations Report and Spreadsheet published

All dates are subject to change

PRMR Schedule

Nov. 10: PRMR Committee Member Education Meeting

Dec. 15: PRMR PAC/LTC* Measure Preview

Mid-January: PRMR Recommendations Meetings

Feb. 25: PRMR Final Recommendations Report published

November | **December** | **January** | **February**

Nov. 12: PRMR Patient Education Meeting

Dec. 16: PRMR Clinician Measure Preview

Dec. 17: PRMR Hospital Measure Preview

*Post-Acute Care/Long-Term Care
All dates are subject to change

Partnership for Quality Measurement
Powered by Battelle

PRMR/MSR Resources for Committee Members

Dr. Meridith Eastman | Battelle

Kate Buchanan | Battelle

Website and Workspace Tour

Dr. Meridith Eastman | Battelle

Understanding CMS Programs

Pay-for-reporting

Provides financial incentives for submitting required quality data, regardless of performance outcomes

Reporting entities may experience a 2-5% point reduction in annual payment updates if they fail to submit required data

Pay-for-performance/ value-based purchasing

Links financial incentives directly to achievement of defined quality, clinical, or operational outcomes

Understanding CMS Program *(cont., 1)*

Hospital Pay-for-Reporting Programs*

- Ambulatory Surgical Center Quality Reporting Program
- Hospital Inpatient Quality Reporting Program
- Hospital Outpatient Quality Reporting Program
- Inpatient Psychiatric Facility Quality Reporting Program
- Rural Emergency Hospital Quality Reporting Program

PAC/LTC Pay-for-Reporting Programs

- Hospice Quality Reporting Program
- Home Health Quality Reporting Program
- Inpatient Rehabilitation Facility Quality Reporting Program
- Long-Term Care Hospital Quality Reporting Program
- Skilled Nursing Facility Quality Reporting Program

Clinician Value-Based Purchasing Programs

- Medicare Parts C Star Ratings (including contracts that offer Part D)
- Medicare Shared Savings Program
- Merit-based Incentive Payment System

PAC/LTC Value-Based Purchasing Programs

- Skilled Nursing Facility Value-Based Purchasing Program

Hospital Value-Based Purchasing Programs

- End-Stage Renal Disease Quality Incentive Program
- Hospital-Acquired Condition Reduction Program
- Hospital Readmissions Reduction Program
- Hospital Value-Based Purchasing Program
- Medicare Promoting Interoperability Program

*Prospective Payment System-Exempt Cancer Hospital Quality Reporting Program is a Hospital Quality Reporting Program with no payment attached

Decile Table Guidance: Performance

Average score of **all** facilities on the measure (for this measure, a lower score is better)

For example, if the average performance of Decile 3 (0.95%) is considered a plausible, achievable score and the entities in Deciles 4 through 10 improved to reach that score, the estimated number of eligible patients who left without being seen would go down by about 1.6%.

Highest Performers ← **Lowest Performers**

	Overall	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Average Score (Standard Deviation)	1.99 (2.63)	0	0	0.95	1.00	1.00	1.80	2.00	2.77	3.62	6.77
Entities	3,860	386	386	386	386	386	386	386	386	386	386
Patients	136,851,910	6,497,824	6,086,507	6,740,161	23,209,776	10,515,795	10,429,301	21,249,856	16,847,585	16,114,181	19,160,924

Table 1. Importance (Decile by Measure Score, PY2023)

Total number of facilities included in performance data

Each decile contains 10% of the total number of facilities that report on the measure

For example, 386 facilities had an average score of one patient who left the ED without being seen

 Example decile table from 2025 MSR measure 00410-01-C-HOQR - Left Without Being Seen

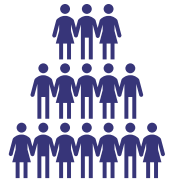
Decile Table Guidance: Reliability

Table 2. Reliability (Decile by Denominator – Target Population Size)

	Overall	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Mean Target Population Size	35,454	1,744	5,528	9,720	14,755	21,296	28,804	37,528	48,862	65,903	120,399
Mean Reliability	99.6	97.1	99.6	99.8	99.9	99.9	99.9	99.9	100.0	100.0	100.0
Entities	3,860	386	386	386	386	386	386	386	386	386	386
Total Patients	136,851,910	673,333	2,133,684	3,751,946	5,695,449	8,220,146	11,118,481	14,485,635	18,860,684	25,438,648	46,473,904

This table sorts entities by the number of patients and reports average reliability along with the number of entities and average number and total patients for each decile. This table can be used to assess the impact of population size on the reliability of an entity's measure score.

Population size can impact reliability estimates



More stable & consistent measure scores

Greater variation

Example decile table from 2025 MSR measure 00410-01-C-HOQR - Left Without Being Seen

Decile Table Guidance: Reliability *(cont.)*

Table 3. Mean Reliability (By Reliability Decile)

Mean	SD	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	IQR
99.6	2.42	97.1	99.6	99.8	99.9	99.9	99.9	99.9	100.0	100.0	100.0	0.2

This table sorts entities by reliability and reports the average reliability by decile. The table also includes the mean, standard deviation, and interquartile range (IQR). This table can be used to see the distribution of the reliability of the entities. A measure is generally considered reliable when the reliability for at least 70% of the individual entities is above 60%.

How to interpret this table: The mean reliability was 99.6 and all entities had a reliability score higher than 60%. This means that the measure can reliably tell the difference between those who are performing better or worse, making it a useful tool for comparing quality of care.

 Example decile table from 2025 MSR measure 00410-01-C-HOQR - Left Without Being Seen

Video Resources

Now Available:

- Navigating the PRMR Tab-Based Measure Display
- How to Submit a Public Comment on MUCs

Coming Soon:

- PRMR and MSR Processes at-a-Glance
- Navigating the MSR Tab-Based Measure Display
- How to Submit a PIE Form for PRMR and MSR

Battelle continues to develop video resources for committee members.

Partnership for Quality Measurement
Powered by Battelle

Day 2 Reflections

Dr. Meridith Eastman | Battelle

Partnership for
Quality Measurement
Powered by **Battelle**