Partnership for Quality Measurement
Powered by Battelle

Measure Developer Workshop Exploring the Science Behind the Scores: Methods, Cost, and Impact

Day 2: Welcome and Introductions

Brenna Rabel | Battelle

October 16, 2025
12:00PM-12:05PM (ET)

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.

Day 1 Recap

- The following topics were discussed yesterday:
  - **Cost and Burden of Quality Measurement** – Guest speaker Dr. Donald E. Casey Jr. discussed how measurement science and economic evaluation can identify high-value improvement efforts and address challenges in scaling interventions.
  - **Performance Gap: Estimating Impact with Heuristics** – This session was on using heuristics, causal reasoning, and empirical methods to assess the actionability and significance of performance gaps.
  - **Presenting and Evaluating Cost Measures** – This session covered the complexities of cost measurement, interpretation in the context of quality care, and application of PQM evaluation criteria.
- Recordings will be posted to the E&M Resources webpage by the end of October.

Agenda

Day 2 – October 16, 12:00-3:30 PM

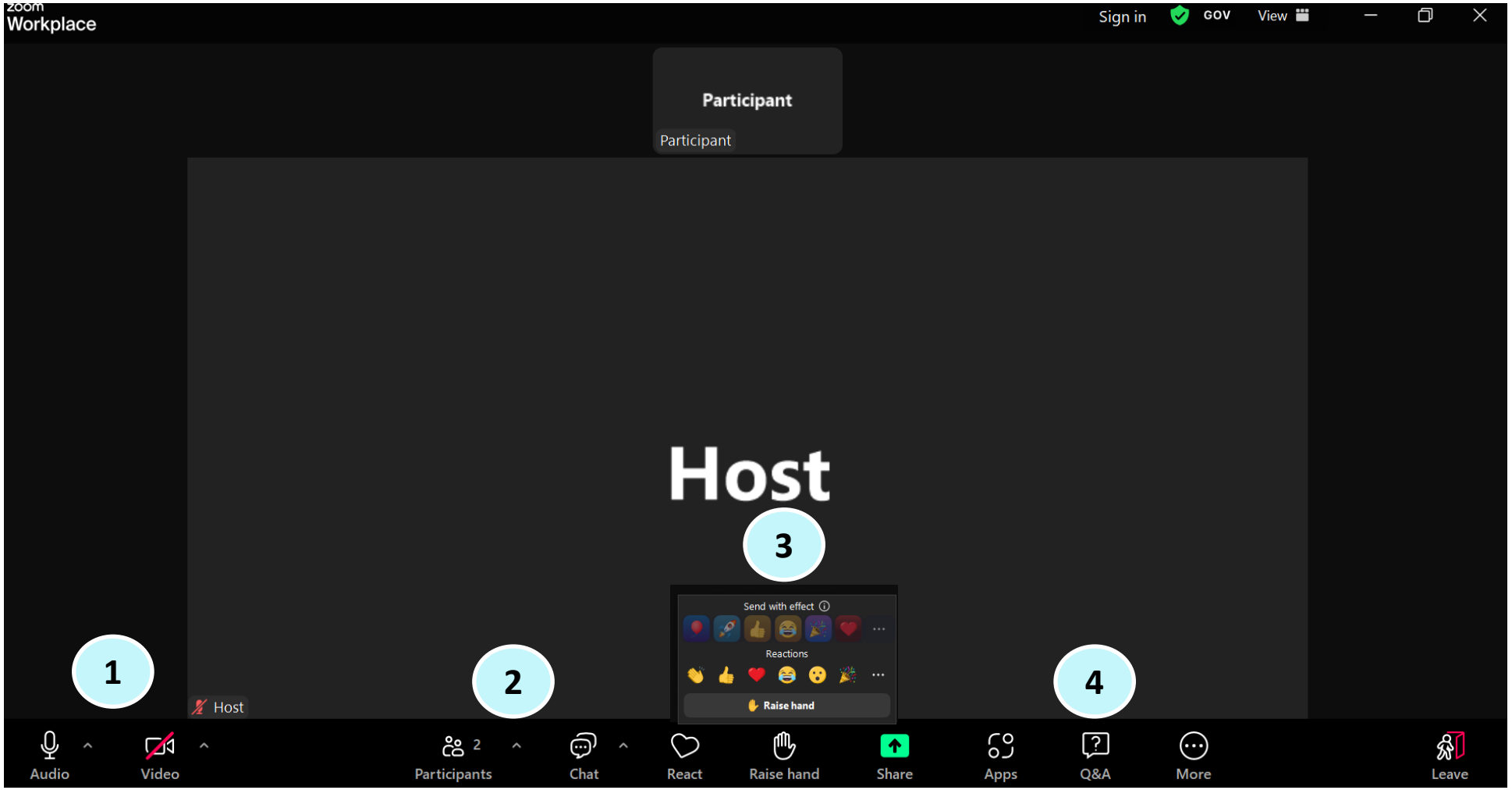
- 12:00 PM: Welcome
- 12:05 PM: Applying Reliability Methods
- 1:20 PM: Validity Best Practices
- 2:20 PM: Break
- 2:30 PM: Risk Adjustment: Variability and Innovation*
- 3:25 PM: Closing

**Includes opportunities for developer feedback to inform future guidance and standards.*
All times are listed in Eastern Time (ET).

Housekeeping Reminders

- Each session will have dedicated time for Q&A.
  - Please include questions in the Q&A box, and Battelle staff will triage at the end of each session.
- The system will allow you to mute/unmute yourself and turn your video on/off.
- The sessions are being recorded and will be posted to the E&M Resources webpage by the end of October.
- If you are experiencing technical issues, please contact the project team via chat on the virtual platform or at PQMsupport@battelle.org.
- We value your input—please look out for a feedback survey on the workshop's content and structure, which will be sent out within 1 week.

Using the Zoom Platform



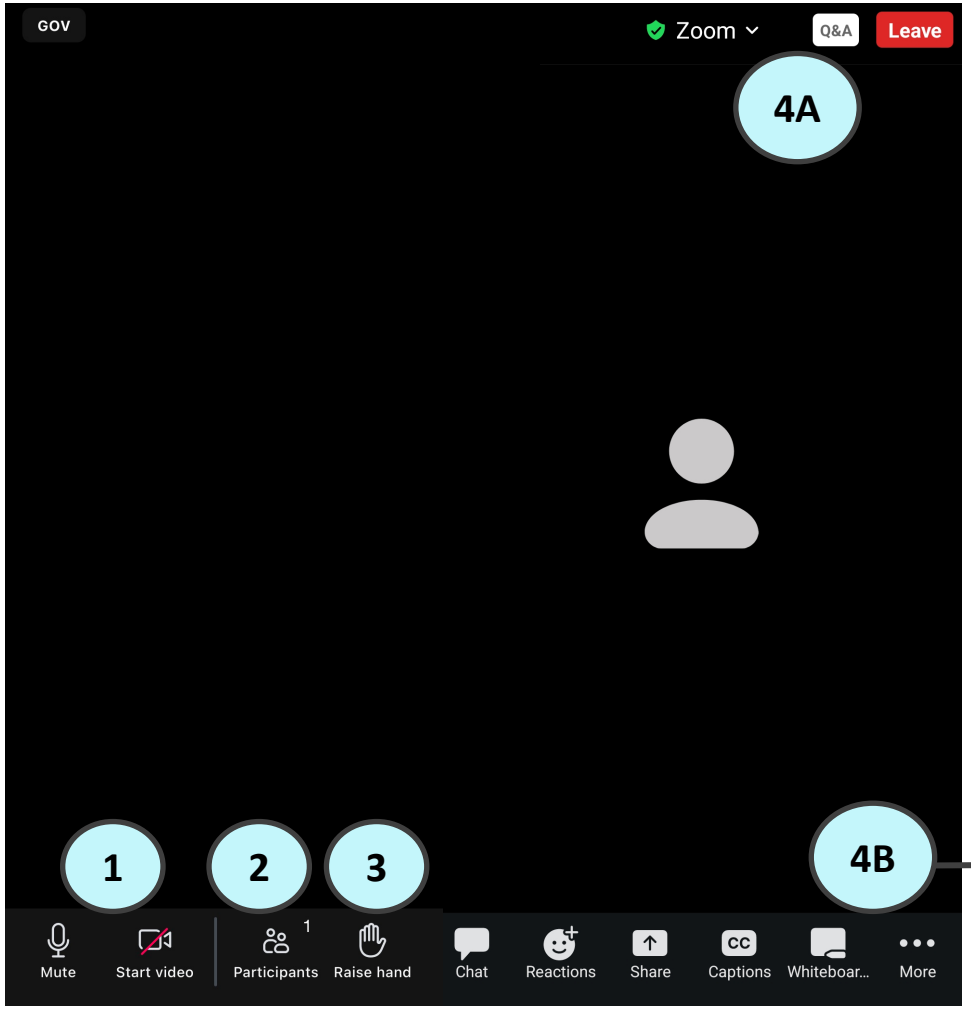
1. Click the lower part of your screen to mute/unmute or start or pause video.

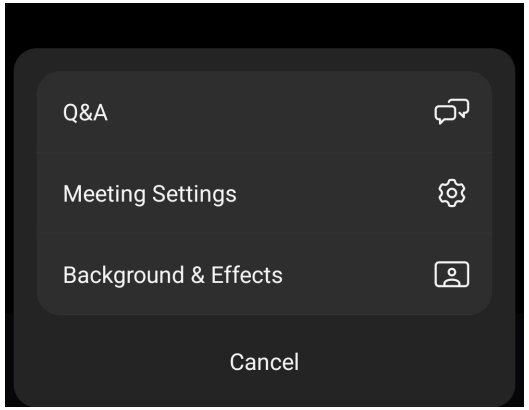
2. Click on the participant or chat button to access the full participant list or the chat box.

3. To raise your hand, select the raise hand button under the react tab.

4. To ask a question, use the Q&A button.

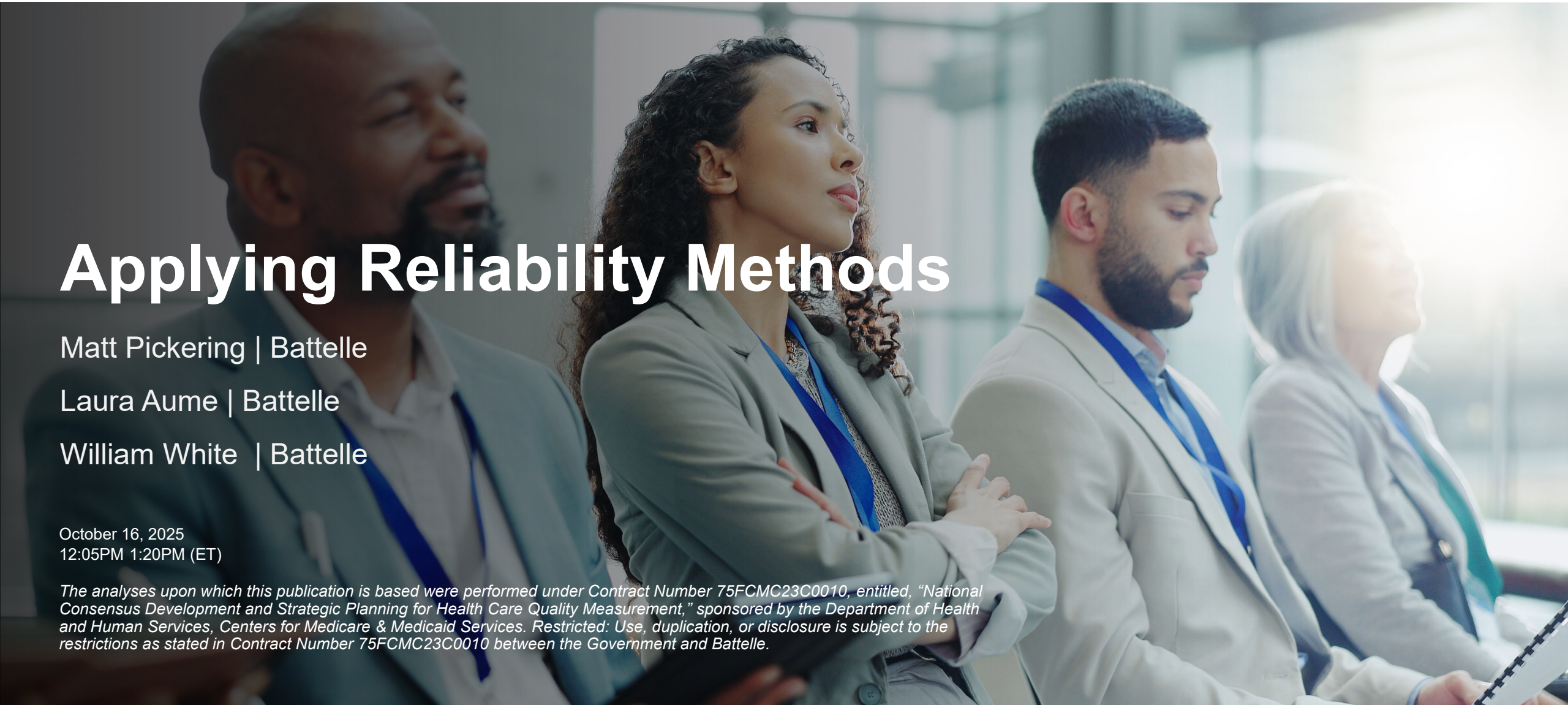
Using the Zoom Platform (Mobile View)





1. Click the lower part of your screen to mute/unmute or start or pause video.
2. Click on the participant button to access the full participant list.
3. To raise your hand, select the raised hand function under the reactions tab.
4. To ask a question, use the Q&A button at the top right of your screen (4A). If you do not see this, you can select the “more” icon (4B).

Partnership for Quality Measurement
Powered by Battelle



Applying Reliability Methods

Matt Pickering | Battelle

Laura Aume | Battelle

William White | Battelle

October 16, 2025
12:05PM 1:20PM (ET)

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.

Meet the Presenters

Matthew Pickering | E&M Task Lead

- Oversees E&M processes and activities
- 10+ years' quality experience

Laura Aume | Data Scientist

- Evaluates measure reliability for E&M
- 30+ years' quality experience

William White | Data Scientist

- Evaluates measure reliability for E&M
- 2+ years' quality experience
- 8+ years' experience in applied statistical analytics and research

Session Objectives and Agenda

- Objectives:
  - Review consensus-based entity (CBE) guidance on reliability methodologies for assessing quality measures, focusing on when and how each method should be applied based on data type and use case.
  - Discuss common challenges and mitigation strategies and trade-offs for entities with low reliability.
- Agenda:
  - Reliability in the Context of Quality Measurement
  - Overview of Reliability Methods at the Person/Encounter and Accountable Entity Levels
  - Challenges in Assessing Reliability
  - Q&A

Reliability in the Context of Quality Measurement

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).

E&M 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.

E&M Testing Requirements

Reliability Testing

- Reliability testing requirements vary depending on whether a measure is being submitted for initial or maintenance endorsement.
- When submitting reliability testing information, measure developers must clearly specify:
  - Level of testing (person/encounter or accountable entity)
  - Method(s) used
  - Statistical results for each test
  - Plain-language interpretation of the results to aid understanding

E&M Testing Requirements

Initial vs. Maintenance Measures

Initial

- Measure developers must demonstrate person/encounter-level reliability for all critical data elements (e.g., those used to calculate the numerator, denominator, and any exclusions).
- Testing is required for manually collected or natural language processing (NLP)-extracted elements (e.g., from clinical notes) but not for structured electronic or claims data if prior reliability evidence exists.
- Existing literature or validation studies may be used to support reliability.
  - The evidence should employ testing approaches identified in E&M Reliability Guidance.
  - Developers must clearly state whether all critical elements are covered by the evidence.

Maintenance

- Measure developers must submit empirical testing at the accountable entity level.
- Reliability is examined across the distribution of accountable entities.
- The associated thresholds are applied to the accountable entity (e.g., facility, clinician, health plan), not the mean or median across entities.

E&M Testing Requirements

Maintenance-Distribution of Reliability Estimates

- Table 2 contains deciles sorted by entity size which can help determine if reliability is impacted by entity size, which may require further exploration.
- Developers can systematically organize and analyze the performance scores and reliability values across different deciles.
- In cases where deciles sorted by reliability differ from those sorted by size an additional table is needed to show the distribution of the entity-level reliability estimates.

Accountable Entity-Level Reliability Testing Results

(Full Measure Submission Form-Table 2)

	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability	0.65	0.69	0.74	0.50	0.70	0.46	0.50	0.62	0.67	0.68	0.74	0.88	0.93
Mean Performance Score	0.17	0.06	0.05	0.13	0.09	0.17	0.21	0.23	0.19	0.21	0.21	0.16	0.09
n of Entities	50	8	5	5	5	5	5	5	5	5	5	5	1
n of Persons/ Encounters/ Episodes	2,500	88	55	63	88	105	147	205	247	308	438	844	201

Step-by-step guidance for calculating Table 2 can be found in our Reliability Guidance document.

	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability	0.65	0.29	0.34	0.43	0.47	0.53	0.60	0.67	0.72	0.80	0.81	1.0	1.0

Overview of Reliability Methods

Types of Data

- To appropriately select statistical tests for reliability, consider the type of data collected at the person/encounter level and how the measure score (accountable entity level) is constructed (i.e., observed rate, risk-adjusted rate, composite).

Categorical data are qualitative and consist of distinct groups or labels	• Nominal, meaning no inherent order (e.g., blood type, gender) or ordinal, meaning ordered categories with meaningful but uneven intervals (e.g., education level). • Can also be classified by the number of response options: binary, meaning two categories (e.g., yes/no) or multinomial, meaning three or more (e.g., eye color, marital status).
Numerical data are quantitative	• Discrete, meaning they represent countable whole numbers (e.g., number of children) or continuous, meaning they can take on any value within a specified range and are measurable (e.g., height, weight, temperature).

Tests of Reliability: Person or Encounter Level

Tests of Reliability: Person or Encounter Level

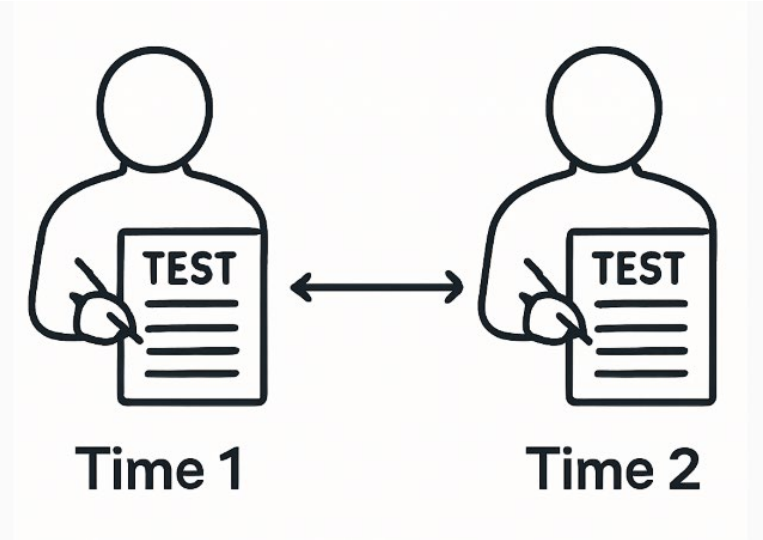
Test-Retest Reliability

- Evaluates the consistency of results when the same individual completes the same instrument or measure at two time points.
- Appropriate when there is a rationale for expecting stability, rather than change, over a short to medium time interval.
- **Common Methods:**

Spearman's Rank Correlation (SR)
Use for ordinal data

Pearson Correlation Coefficient
Use for continuous data that are normally distributed and when a linear relationship is expected

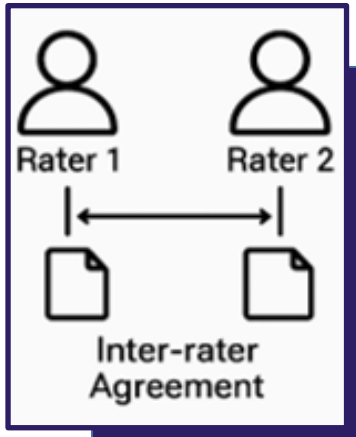
Intraclass Correlation Coefficient (ICC)
Use for continuous or ordinal data
Assesses measurement consistency by randomly splitting patient data within each entity and comparing reliability across subsets ("split-half")



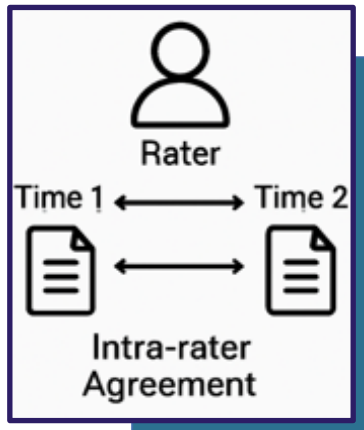
- These methods produce values on a 0-1 scale. A higher correlation value indicates greater test-retest reliability.
- **CBE-Recommended Value for Endorsement:** ≥ 0.5

Tests of Reliability: Person or Encounter Level

Inter- and Intra-Rater Agreement



Inter-rater agreement assesses the extent to which two or more individuals provide consistent judgments or ratings when evaluating the same information.



Intra-rater agreement assesses the consistency of scores assigned by the same individual across two or more occasions.

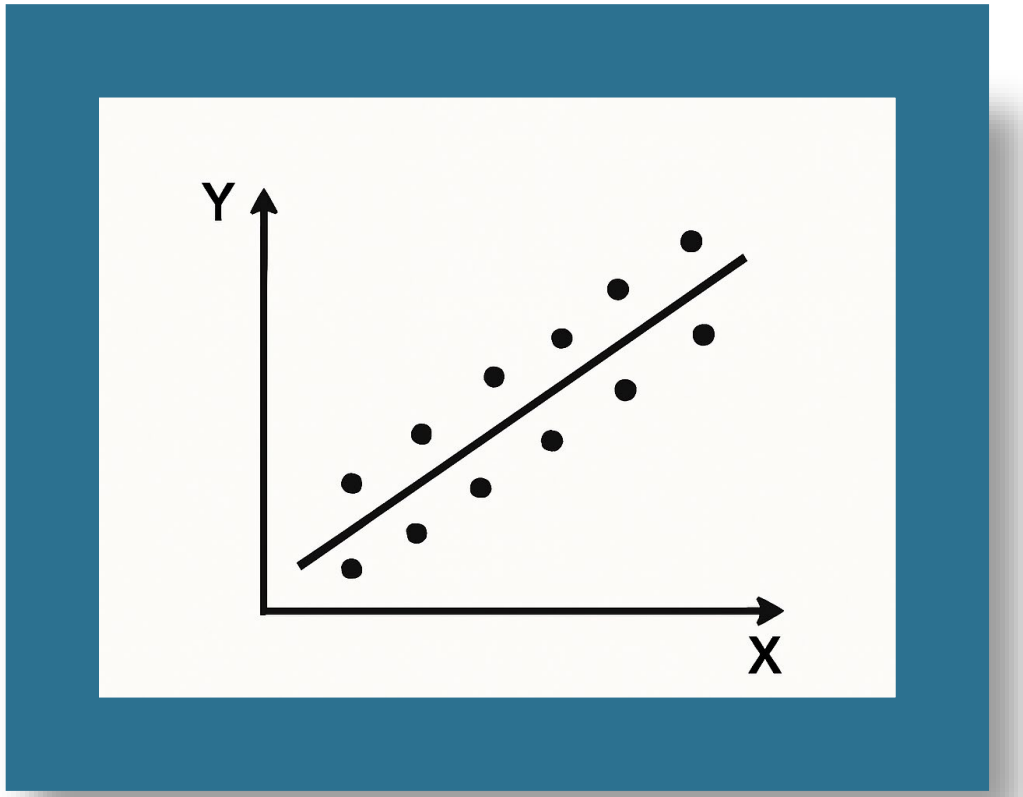
Common Methods:

- Cohen's Kappa: Appropriate for categorical data (e.g., yes/no responses). Adjusts for agreement that could occur by chance.
- Gwet's AC1: A chance-corrected agreement statistic. Well-suited for binary or categorical data that may be imbalanced.
- Kendall's Tau: A non-parametric correlation coefficient. Used with ordinal (ranked) data.
- Percent Agreement: The number of ratings that agree divided by the total number of ratings. Easy to calculate but may overestimate reliability because it does not account for chance agreement.
- **CBE-Recommended Value for Endorsement:** ≥ 0.4

Tests of Reliability: Person or Encounter Level

Linear Relationship

- Linear relationship examines whether person- or encounter-level data from repeated measurements track together in a predictable, linear way.
- **Common Methods:**
  - Pearson Correlation Coefficient: Measures the strength and direction of a linear relationship between two continuous variables.
    - Can be misleading in a reliability analyses because it assesses association rather than agreement.
- **CBE-Recommended Value for Endorsement:** ≥ 0.6



Tests of Reliability: Person or Encounter Level

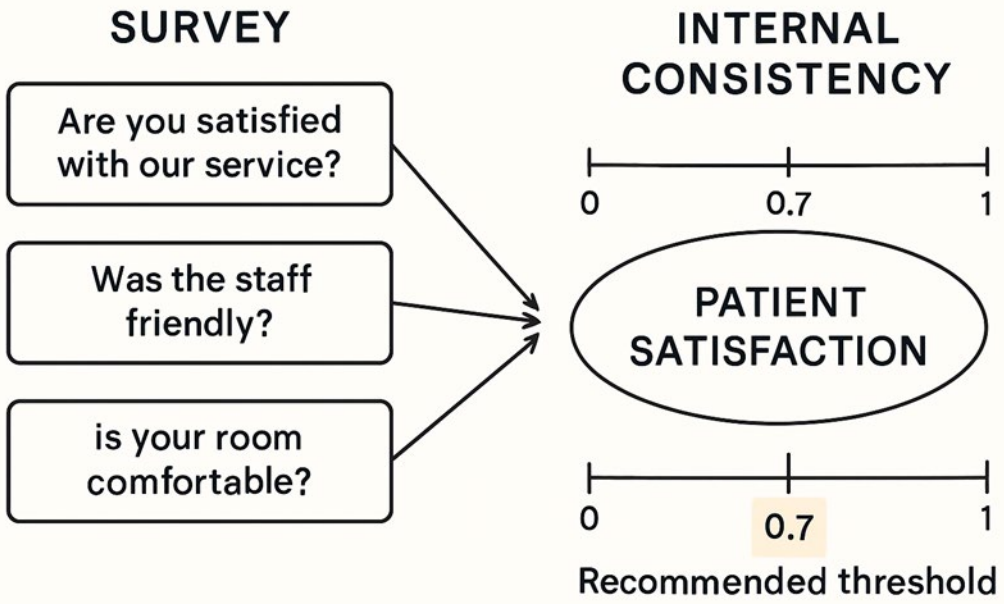
Internal Consistency

- Often used to evaluate the reliability of an instrument.
- Evaluates the degree to which items correlate with one another and reliably capture the same underlying construct (e.g., patient satisfaction).
- **Common Methods:**

Cronbach's alpha (α)
For continuous/ ordinal data

Kuder-Richardson Formula 20 (KR-20)
For binary (yes/no) items

- Both methods yield values from 0 to 1; higher values = greater consistency.
- **CBE Recommended Value for Endorsement:** ≥ 0.7

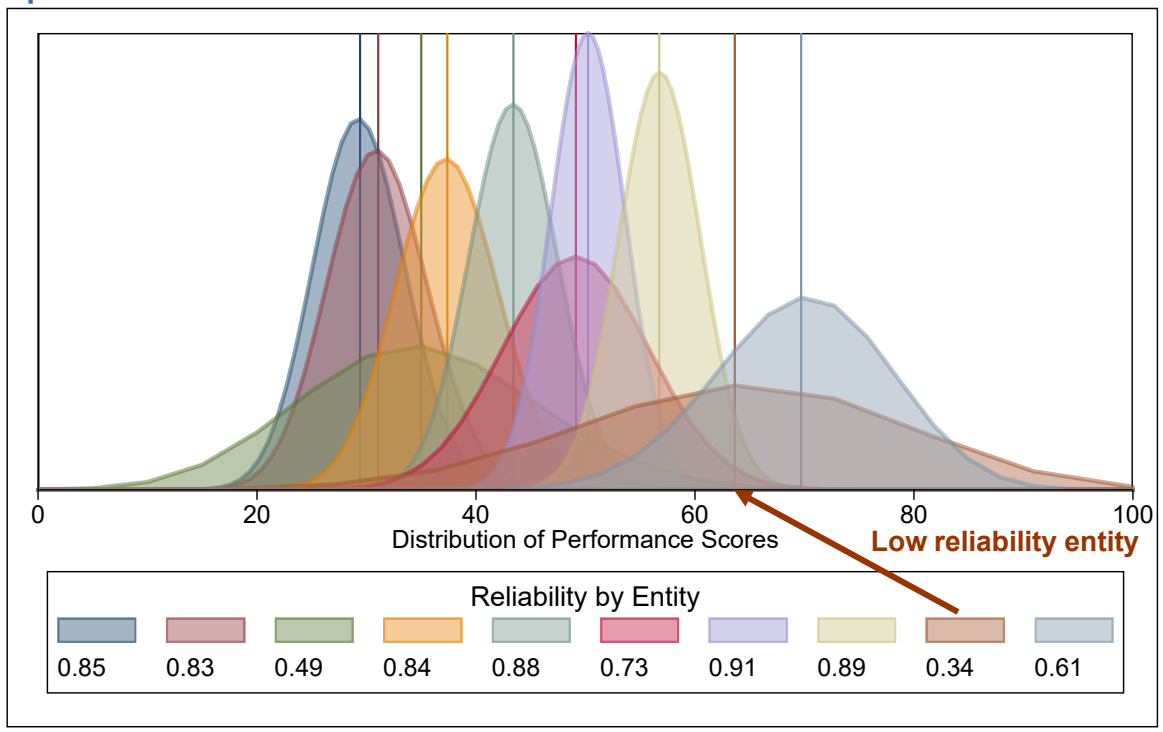


Tests of Reliability: Accountable Entity Level

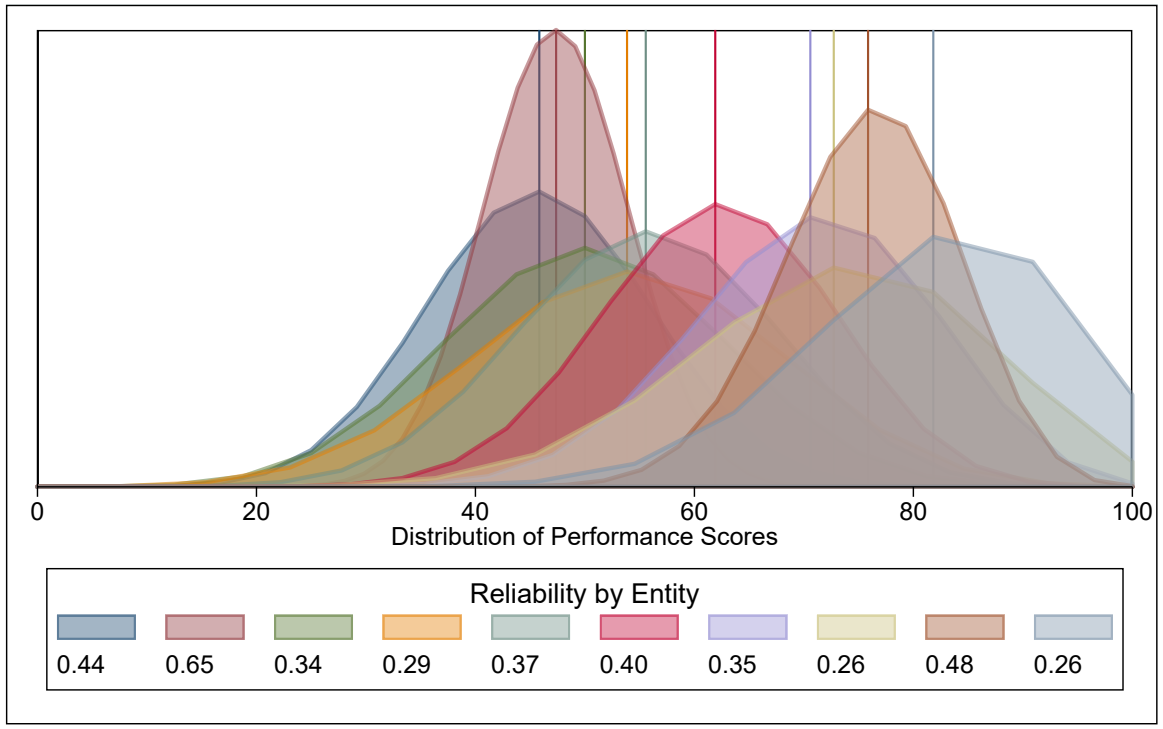
Entity-Level Reliability

- 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



Tests of Reliability: Accountable Entity Level

Signal-to-Noise for Binomial Data

- Signal-to-noise is used to estimate how much of the overall variation in entity-level measure scores is due to systematic differences in performance between entities (the "signal") as opposed to random variation or measurement error ("the noise").

Common Methods

Adams (2009): Estimates signal-to-noise by using beta-binomial parameter estimates for each entity.

Nieser and Harris (2024): Estimates signal-to-noise based on the sample size and the beta-binomial parameter estimates.

We recommend that developers use this approach for binomial measures.

Variations

Interunit Reliability (IUR): Uses total and pooled between-entity variances. Entity level reliability is estimated with the Spearman-Brown prophecy formula. **Developers may also choose to use this approach for binomial measures.**

Empirical Bayes: "Borrows strength" from entities with larger patient or case volume to improve estimates of reliability for entities with smaller patient or case volume.

Logistic Regression: Commonly used for risk-adjusted measures with binomial data.

- **CBE-Recommended Value for Endorsement:** $\geq$ 0.6 for 70% or more of the accountable entities.

Tests of Reliability: Accountable Entity Level

Split-Half (ICC)

- The ICC measures correlation at the patient or encounter level. To assess accountable entity-level reliability, a formula must be applied to obtain an entity-level distribution of performance scores.
- ICC requires an assumption of normality
- Measure developers can apply the Spearman-Brown (SB) prophecy formula to a single ICC value to obtain an entity-level estimate of the ICC based on the entity's sample size. ICC is best suited for measures with continuous data (e.g., average length of hospital stay).
- **CBE-Recommended Value for Endorsement:** ≥ 0.6 for 70% or more of the accountable entities.

$$\hat{r}_i = \frac{k_i \hat{r}}{1 + (k_i - 1)\hat{r}}$$

$\hat{r}$ is the overall reliability estimate
$\hat{r}_i$ is the entity-level reliability estimate
k_i is the ratio of the entity-level sample size to a measure of central tendency (such as mean or median) of the sample size across all entities

Tests of Reliability: Accountable Entity Level

Split-Half (Spearman Rank)

- For Spearman rank (SR), each half of the measure for all entities is sorted and ranked and a correlation is calculated on the ranks.
- Typically for measures based on ordinal data that are not continuous or normally distributed (e.g., a survey measures that uses a 3-point Likert scale of poor, fair, or good).
- The Spearman-Brown (SB) Prophecy Formula can also be applied to a Spearman rank (SR) correlation coefficient to calculate entity-level reliability based on the size of each entity.
- **Methods:** Developers can also use the following methods to improve reliability estimates:

Bootstrap resampling:

The reliability estimate for each entity is calculated as the average Spearman rank correlation.

Permutation Resampling:

The reliability estimate for each entity is the average Spearman rank correlation, with the Spearman-Brown formula applied across all permutations.

We recommend that developers use this approach for continuous measures. If data are normally distributed, developers may choose to apply this process to the ICC.

- **CBE-Recommended Value for Endorsement:** ≥ 0.6 for 70% or more of the accountable entities.

Challenges in Assessing Reliability

Reliability Challenges, Mitigation Strategies, & Tradeoffs

Small Sample Size/Case Volume

Challenge	Mitigation Strategy	Tradeoffs	Evaluation Questions to Consider
Small Sample Size/Case Volume	Aggregate data across multiple years to increase sample size.	• Slower feedback to accountable entities. • Data may be outdated and less actionable.	• Are measure results timely enough for the intended purpose? • Are measure results consistent across years, suggesting slower feedback may be acceptable? • Do users prefer greater stability over rapid feedback?
	Combine individual measures into a composite.	• Loss of detail for individual measure areas. • May obscure poor performance in specific areas.	• Does the composite maintain conceptual coherence? • Is the sum better than the individual components? • Are individual component results available for review? • Are there processes to address poor performance in specific areas?

Reliability Challenges, Mitigation Strategies, & Tradeoffs

Some Low-Volume Entities

Challenge	Mitigation Strategy	Tradeoffs	Evaluation Questions to Consider
Some Low-Volume Entities	Use hierarchical modeling to "shrink" extreme values toward the mean.	• Adds complexity. • May mask the performance of outliers (very high or low performers).	• Is there transparency in how entities are adjusted? • Do the reliability estimates of high and low performers change over subsequent measurement periods? • Does this approach preserve meaningful performance differences? • Are users comfortable interpreting adjusted scores?
	Report at a higher level to combine data across entities (e.g., system level).	• Loss of granularity. • May weaken local accountability.	• Will users still find results meaningful and actionable? • Can entities review their results internally for more specific insights? • Is local accountability still possible with aggregated reporting?
	Exclude low-volume providers by setting a minimum threshold.	• Excludes small providers. • May omit safety-net or underserved populations.	• Is a low reliability measure better than no measure at all due to its importance and no alternative measure? • What percentage and types of entities are excluded? • Are excluded providers serving vulnerable populations? • Are alternative monitoring or quality improvement strategies in place (e.g., confidential feedback) for excluded entities?

Reliability Challenges, Mitigation Strategies, & Tradeoffs

Infrequent Events or Rare Outcomes

Challenge	Mitigation Strategy	Tradeoffs	Evaluation Questions to Consider
Infrequent Events or Rare Outcomes	Combine similar outcomes or events into one measure.	• Loss of specificity. • May conflate clinically distinct issues.	• Does combining events still reflect meaningful performance? • Are stakeholders aligned on what is appropriate to group? • Can disaggregated data still be analyzed for root cause analysis?
	Extend measurement period (e.g., over multiple years).	• Delays event detection. • Potential increased risk from slower signal recognition.	• Is delayed feedback acceptable for this event type? • Are risks of delayed detection outweighed by improved reliability? • Can other mechanisms provide early warning signals?

Reliability Challenges, Mitigation Strategies, & Tradeoffs

Little Variation Between Entities

Challenge	Mitigation Strategy	Tradeoffs	Evaluation Questions to Consider
Little Variation Between Entities	Increase measure sensitivity (e.g., use more granular categories or continuous scoring).	• Increased complexity may reduce interpretability.	• Will users understand and accept more sensitive but complex measures? • Are there subpopulations where variation is present?

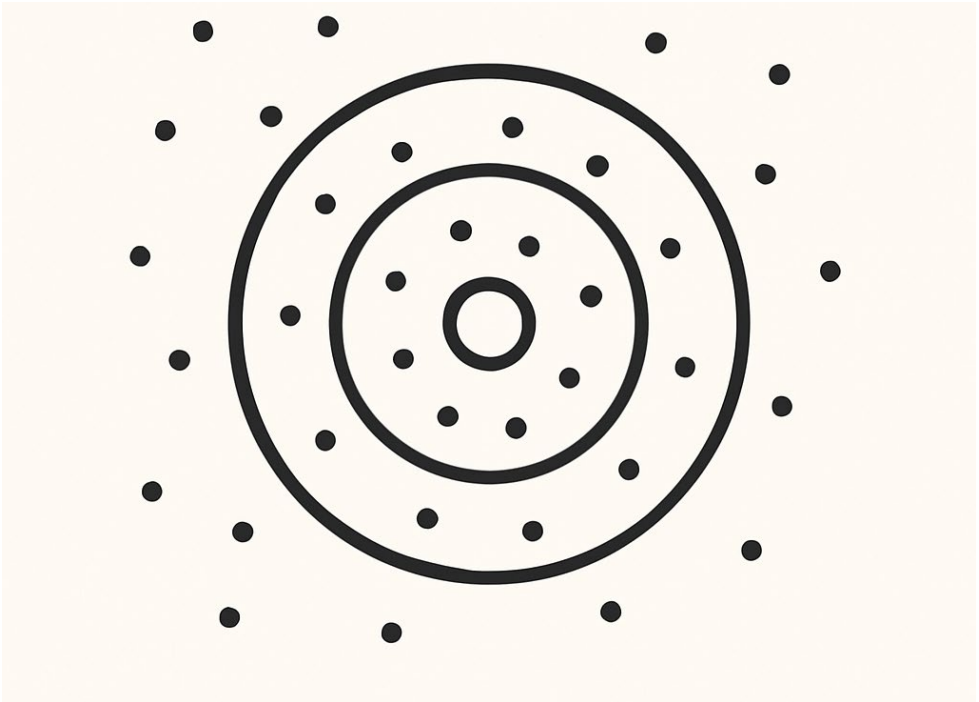
Reliability Challenges, Mitigation Strategies, & Tradeoffs

Measure is Topped Out

Challenge	Mitigation Strategy	Tradeoffs	Evaluation Questions to Consider
Measure is Topped Out	Determine if the performance standard or threshold should be raised.	• Demotivates entities if new targets are perceived as unattainable.	• Are there higher standards or new evidence that justify a more challenging measure?

When is Low Reliability Acceptable?

- Low reliability does not necessarily mean a measure is not useful.
  - The measure's utility depends on available alternatives and the consequences of its use.
  - When the measure focus is closer to the continuum of "should not happen," (e.g., never events) less emphasis may be placed on reliability.
- As a best practice, the submission should describe the decision-making process.
  - Include any input received from interested parties (e.g., technical expert panels, public comment) that informed the final determination.



When is Low Reliability Acceptable?

Scenarios Where Low Reliability May Be Acceptable

Scenario in Which Low Reliability May Be Acceptable	Description	Example
Direct Causal Relationship	The measure focuses on an outcome with a clear, direct causal link to the intervention, with no alternative mechanisms or confounding person-level factors.	A measure of surgical site infections following a specific procedure, where the procedure itself is the only plausible cause.
No Alternative for Decision-Making	When no other reliable measures are available, using a low reliability measure may still provide valuable information for decision-making.	A rural health program uses a low reliability measure of patient satisfaction because no other patient-reported data is collected in the area.
Availability of a Complementary High-Reliability Measure	A related structure, process, or outcome measure with high reliability is available to support interpretation and use of the lower reliability measure.	A measure of complication rates (low reliability) is paired with a measure of adherence to surgical protocols (high reliability).
Mitigation Would Reduce Validity or Participation	Strategies to improve reliability (e.g., combining data over multiple years) would reduce the measure's validity or discourage provider engagement.	Extending data collection to 3 years would increase reliability but obscure recent improvements and disincentivize participation in an accountability application.

When is Low Reliability Acceptable?

Scenarios Where Low Reliability May NOT Be Acceptable

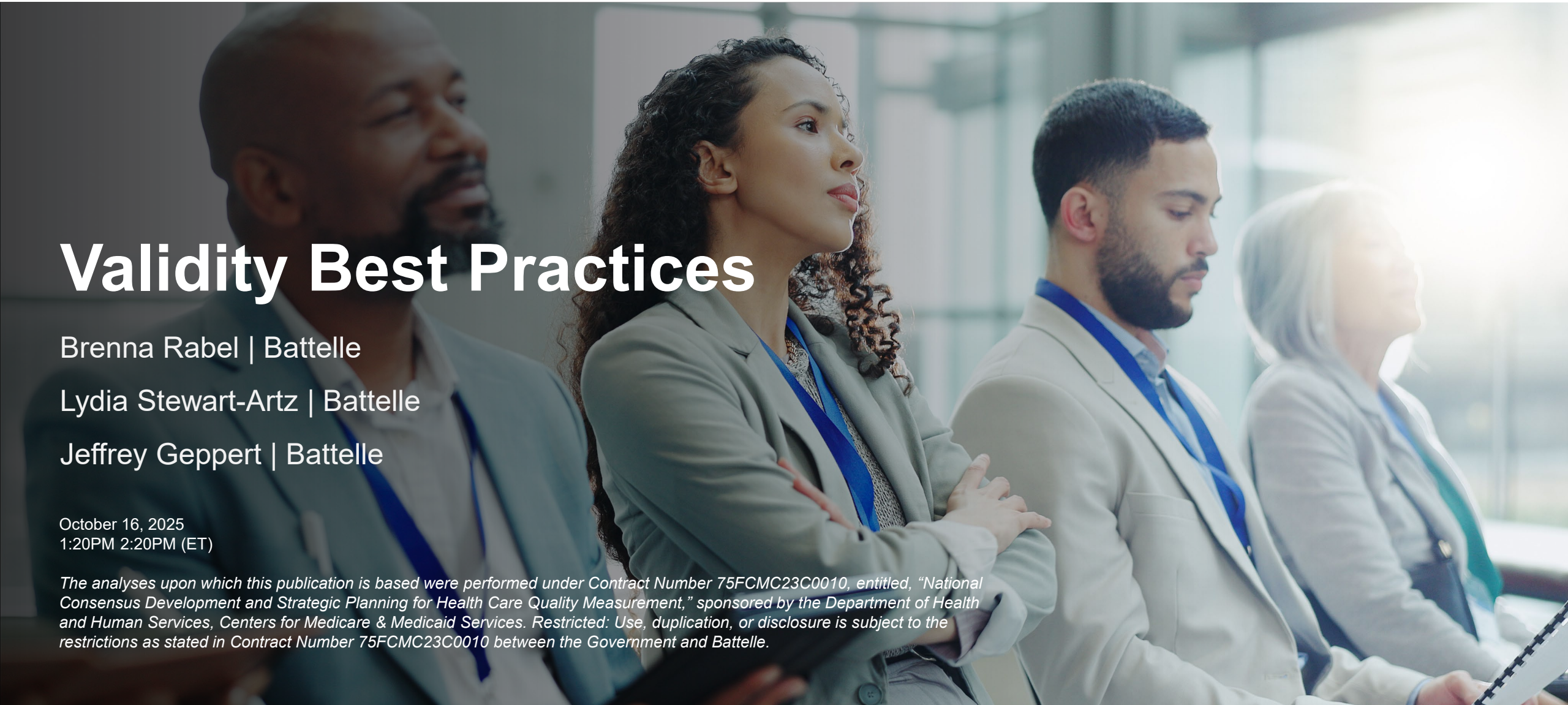
Scenario in Which Low Reliability Is Not Acceptable	Description	Example
Disproportionate Impact on Vulnerable Populations	When specific, identifiable populations are more likely to receive care from providers with low reliability scores, leading to potential biases.	Low reliability in accountable entities serving marginalized communities (e.g., safety-net hospitals) may result in unfair penalties that disproportionately affect underserved populations.
Potential for Material Harm	When decisions informed by the measure could result in serious consequences for providers or patients, such as public reporting, payment adjustments, or loss of trust.	A low reliability mortality rate measure used in a payment program could lead to unjust financial penalties for accountable entities without accurately reflecting performance.
Risk of Waste, Futility, or Injustice	When low reliability compromises the measure's overall value (e.g., reducing its importance, validity, or usability) leading to wasted resources or unjust outcomes.	A low reliability measure of provider communication is used to rank clinicians, despite weak correlation with actual care quality, resulting in misleading comparisons and reputational harm.

Resources

- The Blueprint Measure Lifecyle content on the **Measures Management System (MMS) Hub** provides general guidance about the types, uses, and tests for measuring **reliability**.
- **The E&M Reliability Guidance** provides measure developers with the tools to ensure their measures are robust and scientifically sound by:
  - Effectively applying reliability methods
  - Interpreting results in preparation for endorsement consideration

Questions & Answers

Partnership for Quality Measurement
Powered by Battelle



Validity Best Practices

Brenna Rabel | Battelle

Lydia Stewart-Artz | Battelle

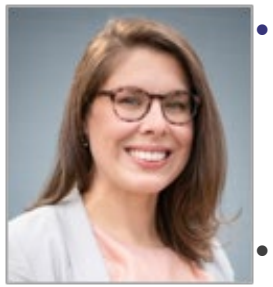
Jeffrey Geppert | Battelle

October 16, 2025
1:20PM 2:20PM (ET)

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.

Meet the Presenters

Brenna Rabel | CBE Technical Director



- Facilitates collaboration across CBE activities to ensure consistency and excellence
- 10+ years' health care, public health, and quality experience

Lydia Stewart-Artz | PRMR/MSR Evaluation Lead

- Oversees PRMR/MSR measure evaluation and committee review
- 5+ years' quality experience

Jeffrey Geppert | Sr. Research Leader

- Leads Measurement Science team for E&M
- 27+ years' measurement science, health care, and quality experience

CBE: consensus-based entity E&M: Endorsement and Maintenance PRMR: Pre-Rulemaking Measure Review MSR: Measure Set Review

Session Objectives and Agenda

- Objectives
  - Review best practices for accountable entity-level validity testing, with a focus on aligning hypothesized measure relationships, interpreting correlation patterns, and leveraging mechanistic evidence to support validity consistent with CBE principles.
- Agenda
  - Background and Discussion of PQM Validity Rubric and Articulating Hypothesized Relationships
  - Assessing and Interpreting Expected Correlation Direction and Strengths
  - Causal and Mechanistic Explanations and Validity Best Practices
  - Q&A

Background and Discussion of PQM Validity Rubric & Articulating Hypothesized Relationships

What is Validity?

Validity is “an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy of appropriate interpretations and actions on the basis of [the measure].”
– Messick (1989); Standards for Educational and Psychological Testing (2014)

- Judgment – validity is a human judgment balancing benefits and harms
- Degree – validity is based on an accumulation of evidence and argument with a progressive reduction in residual risk
- Empirical evidence and theoretical rationales – to substantiate validity claims, one needs both theory (plausible argument) and evidence (broadly defined)
- Appropriate interpretations and actions – the causal inference being substantiated in a validity claim

Note: The association of the measure focus with a material outcome supports an Importance claim, rather than a Validity claim.

Why is Measuring Validity Important?

- **In principle, the greater the degree of validity, the . . .**
  - More likely that a person who selects a better-performing entity will experience the measure focus (and, if the measure is "Important," experience a material outcome), and
  - More likely that an entity that chooses to allocate resources will transform to a better-performing entity
- However, this logic is vulnerable to error if based on the assumption that relationships observed at the level of groups (i.e., entity) necessarily hold for individuals within those groups (i.e., persons).
- These best practices are intended in part to better support the logic of value-based payment and comparative reporting.

Updated PQM Criteria for Validity

Person or Encounter Level

- Methodology employed and the analytic approach presented are appropriate and thorough[1]; AND
- Results of empirical testing or prior evidence adequately demonstrates that all critical data elements (numerator, denominator, exclusions) are valid with limited or no threats to validity present.

Accountable Entity Level

- Methodology employed is adequate and the analytic approach presented is appropriate and thorough (i.e., **clear reasoning for conducting a correlation analysis with another quality measure, clear hypothesis for correlations, and supportive evidence from mechanistic studies to justify correlation results**); AND
- Results of empirical testing adequately demonstrate that the measure is valid with limited or no threats to validity; AND
- The interpretation of the empirical results supports and interference of validity; AND
- [For initial endorsement] Face validity is adequate.

[1]This includes empirically assessing the impact of missing data and exclusions, as appropriate.

Hypothesized Relationships Between Two Measures

- The most common form of accountable entity validity testing is correlations between the measure of interest and related process and outcome measures.
  - However, correlations between two invalid measures do not make both measures valid.
  - Also, correlations among multiple measures and then selecting meaningful correlations after the fact is not compelling evidence of validity.
- Correlation must be interpreted through the lens of shared mechanisms.
  - The degree of overlap among quality constructs (e.g., mechanism complexes or quality interventions) can be quantified using various methods, providing a basis for interpreting correlations through shared mechanisms.
- Correlation does not imply causation; however, causation does imply correlation.
  - If two measures share 30% of causal factors, then the correlation ought to be about 0.30 (with some potential attenuation due to measurement error).

Hypothesized Relationships Between Two Measures *(cntd., 1)*

- Articulating hypothesized relationships involves:
  - First, examining the relationships between related process and outcome measures based on the overlap in the theoretical quality constructs they are supposed to reflect.
  - Second, examining the empirical correlations to see if they have the expected direction and magnitude.
  - Third, adjust the theory (overlap) if the empirical data are different than expected.
  - **Example**: A developer assesses patient falls measure with another indicator of patient safety, use of patient safety protocols. The developer found a correlation coefficient of -0.25 between use of safety protocols and reduction in patient falls. Although in the expected direction, this correlation may be lower than expected based on the theoretical constructs that both measures aim to improve patient safety. There may be competing causes or counteracting mechanisms.

Assessing and Interpreting Expected Correlation Direction and Strengths

Four Tiers of Correlation Justification

Table 1. Correlation Justifications

Level	Justification	Maturity/Claim	Validity Claim
Low	Descriptive	Level I-weak	A statement that there is an overlap in the quality construct without an explicit articulation of the specific mechanisms in common.
Medium	Qualitative	Level II-moderate	An explicit articulation of the specific mechanisms that overlap in the quality construct and the mechanisms that do not overlap; a corresponding qualitative expectation about the magnitude of the correlation (e.g., moderate overlap implies moderate correlation.
High	Qualitative-relative	Level II-strong	An explicit articulation of the magnitude of the overlap in the quality construct that should correspond to the magnitude of the correlation. The articulation is relative: the overlap between A and B is greater than the overlap between A and C; therefore, the correlation between A and B should be greater than the correlation between A and C.
Highest	Qualitative-absolute	Level III-strong	An explicit articulation of the magnitude of the overlap in the quality construct that should correspond to the magnitude of the correlation. The articulation is absolute: the overlap between A and B is 30%; therefore, the correlation between A and B should be 0.30. Note that there are still 70% of the causal factors unexplained (note: observed correlation may be attenuated by measurement error, even with high-quality construct overlap).

- Table 1 describes how to make this correlation analysis more rigorous.
  - The justifications rank from the lowest to the highest, in terms of less residual risk.
  - The goal is to move from more descriptive and qualitative justifications to more quantitative justifications.
  - There are various methods that may be used to quantify the degree of overlap is causal.

Low Justification

- **Justification:** Descriptive
- **Maturity/Claim:** Level I-weak
- **Defining Characteristics:** Stating that measures are related because they share some conceptual domain but do not detail the mechanisms in common.
- **Example:** “Measure A and measure focus B both reflect quality of care, so we can expect them to be correlated.”
- **Limitations:**
  - Lacks specificity – assuming overlap in the quality construct without proof.
  - Provides minimal support – evidential pluralism demands more than just a statement.

Medium Justification

- **Justification:** Qualitative
- **Maturity/Claim:** Level II-moderate
- **Defining Characteristics:** Identifies overlapping mechanisms between two measures and gives qualitative expectation of correlation strength.
- **Example:** “Because A and B share several key processes (X and Y) but differ in others, we expect a moderate correlation.”
- **Strengths:**
  - Begins to tie to correlation theory.
  - Acknowledges both shared and unshared components.

High Justification

- **Justification:** Quantitative-relative
- **Maturity/Claim:** Level II-strong
- **Defining Characteristics:** The argument becomes quantitative in a relative sense, specifying not only that A and B should correlate, but how that correlation should compare to other known relationships, based on the degree of construct overlap; comparative expectation.
- **Example:** “The overlap between A and B is greater than between A and C; therefore, the correlation (A, B) should be greater than the correlation (A, C).”
- **Strengths:**
  - Sets up a falsifiable, theory-driven prediction.
  - Remains “relative” by not pinning an absolute value to the correlation.

Highest Justification

- **Justification:** Quantitative-absolute
- **Maturity/Claim:** Level III-strong
- **Defining Characteristics:** Provides a specific, numeric prediction of the correlation based on the estimated proportion of shared quality construct.
- **Example:** “Measure A captures 30% of the factors that determine measure focus B; therefore, we expect a correlation of about 0.30 between A and B.”
- **Strengths:**
  - Most rigorous tier.
  - Accounts for measurement error.

Example: CBE #3558 - Initial Opioid Prescribing for Long Duration (IOP-LD)

- Validity: Met
  - The developer evaluated convergent validity, which assesses whether a measure behaves as expected in relation to other conceptually similar measures – e.g., other health plan-level quality measures that focused on reducing potentially inappropriate medication use and improving patient safety.
  - They identified three measures with construct overlap and one without.
    - COB: Concurrent Use of Opioids and Benzodiazepines
    - POLY-ACH: Polypharmacy: Use of Multiple Anticholinergic Medications in Older Adults
    - POLY-CNS: Polypharmacy: Use of Multiple Central Nervous System (CNS)-Active Medications in Older Adults
    - OHD: Use of Opioids at High Dosage in Persons Without Cancer

Example: CBE #3558 - Initial Opioid Prescribing for Long Duration (IOP-LD) *(cntd., 1)*

Measure	Full Name	Hypothesized Correlation with IOP-LD	Rationale
COB	Concurrent Use of Opioids and Benzodiazepines	Positive, weak-to-moderate	Both assess risky opioid-prescribing practices; expected to reflect similar quality and safety strategies at the health plan level.
POLY-ACH	Polypharmacy: Use of Multiple Anticholinergic Medication in Older Adults	Positive, weak-to-moderate	Both target inappropriate prescribing in older adults; overlap in patient safety and medication management interventions.
POLY-CNS	Polypharmacy: Use of Multiple CNS-Active Medications in Older Adults	Positive, weak-to-moderate	Both focus on CNS-related medication risk and polypharmacy, similar safety goals, and prescribing overnight mechanisms.
OHD	Use of Opioids ay High Dosage in Persons Without Cancer	No correlation expected	Different populations (initial vs. chronic use), different prescribing contexts, and different clinical and regulatory levers.

Example: CBE #3558 - Initial Opioid Prescribing for Long Duration (IOP-LD) (*cntd., 2*)

- Expected correlations (with COB, POLY-ACH, and POLY-CNS) were observed and statistically significant.
- No meaningful correlation was observed between IOP-LD and OHD, consistent with hypotheses.
- These results support the criterion validity of the IOP-LD measure:
  - It performs similarly to conceptually related measures.
  - It behaves differently from unrelated measures, confirming it is measuring a distinct concept.
- Empirical support for ruling-in responsible mechanisms also includes several empirical studies and reports (e.g., Point-of-Sale Safety Edits, Concurrent Drug Utilization Review [DUR], Drug Management Programs [DMPs], Clinical Guidelines [e.g., CDC 2022 Guidelines], Educational Outreach and Use of Patient Advisory Panels).

Summary

- Correlation alone is not enough.
  - Simply labeling a correlation as “strong” or “weak” is insufficient for supporting a validity claim.
- Theoretical justification is essential.
  - Correlations must be interpreted considering shared mechanisms between the measure of interest and the related process and/or outcome.
- This approach demonstrates alignment with modern validity theory.
  - This approach aligns with modern validity theory, which emphasizes combining empirical data with theoretical rationale.
- Evidential pluralism is actively applied in the analysis.
  - Table 1 operationalizes evidential pluralism by requiring that correlations be explained through mechanism-based reasoning.

Summary *(cntd., 1)*

- Anticipated and accidental correlations are distinguished in the analysis.
  - A high correlation that is theoretically expected is stronger evidence of validity than one found by chance.
- Structured interpretation is applied to the evaluation of correlation evidence.
  - Table 1 helps set expectations for how different levels of correlation evidence should be justified and interpreted.

Causal and Mechanistic Explanations & Validity Best Practices

Providing Causal Explanation

- If the correlations are not as strong as expected, providing a causal explanation helps in understanding whether the measures are valid or if other factors are influencing the results.
  - **Example**: To explain the low correlation, the developer might investigate whether different implementation levels of safety protocols across various departments affect the effectiveness of these protocols in reducing patient falls. They might also consider external factors such as patient demographics or departmental differences that could influence the results.
- The causal inference being substantiated in a validity claim is that a quality program and the entity response to that program (A) is “causing” the measure focus (B) (A causes B).

Evidence from Mechanistic Studies

Literature Review

- This involves reviewing existing research to see if it supports the constructs the measures are intended to assess. This includes mechanistic studies relevant to the measure's focus. While much literature may pertain to related topics, it is essential to justify the external validity of the evidence concerning the specific measure.

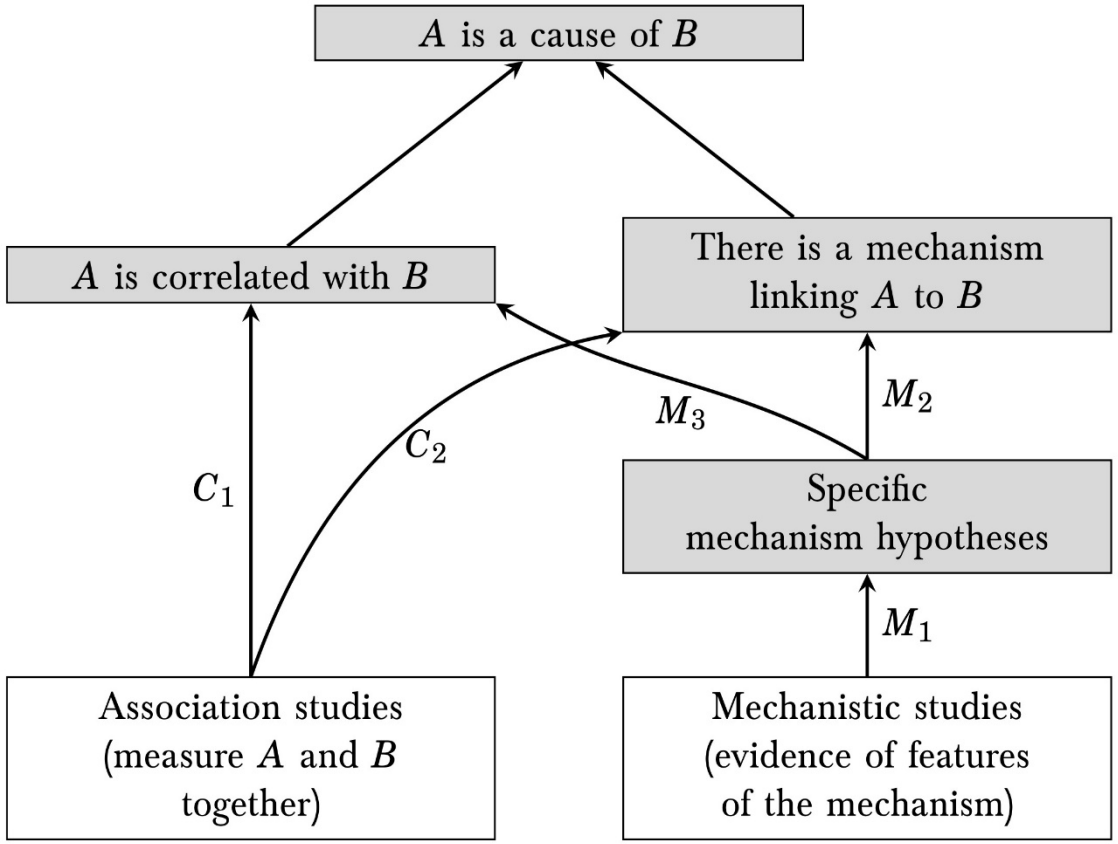
Empirical Analysis

- This involves correlation analyses between aspects of the mechanism and the measure focus across entities (or persons).

Justification of External Validity

- This is crucial when the direct evidence from the literature does not exactly match the constructs of the measures. If the existing literature does not directly address the measure but is deemed relevant, developers must provide a justification for its applicability. This includes detailing how the findings from the literature support the measure's logic model and intended outcomes.

Accountable Entity Validity Claims



- **A is a cause of B**
  - A: quality program and entity response
  - B: measure focus
- **Association claims**
  - A is correlated with B
- **General mechanism claims**
  - A is responsible for B
  - Accounts for the association

Source: Shan, Y., Williamson, J. (2023). Evidential Pluralism in the Social Sciences. United States: Taylor & Francis.

Accountable Entity Validity Claims (*cont., 1*)

Table 1. Substantiating Causal Claims

Claim	Association studies	Mechanism studies
Causal claim (A is a cause of B)	(Prone to bias)	(Prone to complexity)
Correlation claim (A and B are probabilistically dependent conditional on potential confounders C)	Level 1 Test for an association between A and B (C1)	Level 3 Mechanism complex features are **ruled-in** + unsystematic observations (M3) Residual risk: counteracting mechanism
General mechanistic claim (there is a complex of mechanisms that invokes A as partially responsible for B and that can account for the extent of the correlation)	Level 2 Non-causal connections and confounding are **ruled out** (IBE) (C2) Residual risk: competing causes	Level 4 Confirm existence of suitable mechanism (M2)
Specific mechanism hypothesis (posit features of such a mechanism complex)	-	Confirm presence of hypothesized features (M1)

Source: Shan, Y., Williamson, J. (2023). Evidential Pluralism in the Social Sciences. United States: Taylor & Francis.

Note: The symbols C1, M1, etc. refer to the figure on the preceding slide (24).

Residual Risk

- Residual risk refers to unexplained variance that could be due to confounding, bias, or unmeasured factors, threatening the validity of causal claims.
- Table 2 lists rival hypotheses for an observed association between a quality intervention (A) and the measure focus (B) that must be **ruled out** to support a valid claim.
- Addressing these scenarios is essential to ensure that observed associations reflect true causal effects.

Table 2. Other Possible Explanations to Rule-Out

Possible Explanations	Description
Causation	A (quality program & response) is a cause of B (measure focus)
Reverse causation	B is a cause of A
Confounding	C is a common cause of both A and B (risk-adjustment)
Performance bias	A group identified and treated differently than not A group
Detection bias	B is measured differently in A group than in not A group
Chance	Random (reliability-adjustment)
Fishing	Association between A and some B
Temporal trends	A and B change over time for independent reasons
Semantic relationships	A and B have overlapping meaning
Constitutive relationships	A is a component of B
Logical relationships	A and B are logically overlapping
Nomological (law) relationships	Association between A and B due to a natural law
Mathematical relationships	A = B + C

Source: Adapted from Parkkinen *et al.* 2018; Shan & Williamson 2023

- Systematically ruling out these alternatives strengthens the validity argument and aligns with modern validity theory and causal inference principles

Confounders Versus Counteracting Mechanisms

- Both confounders and counteracting mechanisms can weaken the observed relationship between and quality intervention and measure focus but in different ways

Table 2. Distinction: Confounder vs. Counteracting Mechanism

Term	What It Explains	Causal Role	Example
Confounder	A third variable that is associated with both the intervention (or exposure) and the outcome, potentially biasing the estimated effect.	*Competes for causal attribution.*	Patients with poor insight may be more likely to decline long-acting injectables (LAI) and have lower adherence – making it appear that LAIs are less effective.
Counteracting Mechanism	A context-triggered causal process that dampens, negates, or reverses the effect of the primary mechanism.	*Interferes with mechanism activation.*	Side effects may trigger intentional non-adherence, reducing the impact of adherence-promoting interventions.

- Distinguishing between the two helps analysts interpret weaker-than-expected correlations:
  - Confounders require better control (e.g., risk adjustment), while counteracting mechanisms call for deeper understanding of context and intervention design.
- This distinction is crucial for making accurate validity arguments and for deciding whether to adjust study design or refine the intervention.

Validity Best Practices

Test the measure as specified, and, if performing accountable entity-level testing, conduct testing at all levels specified.

Describe the data used for all aspects of test (e.g., reliability, validity, risk adjustment).

Clearly describe the testing approach. Do not just state the type of test being used; explain why the test is being used. Explain what is being tested and any consideration of missing data and how they were addressed.

For face validity, disclose relevant experts and assess if the logic model can be implemented by accountable entities to improve outcomes. Include at least 12 experts, noting their consensus and any disagreements.

When conducting correlation analyses, clearly describe what is hypothesized and why, citing existing evidence accordingly to support the hypothesized relationship.

Clearly describe and summarize results, regardless of if an attachment is provided.

Interpret the results in terms of demonstrating validity for each level and type of validity testing conducted. What are the harms associated with low validity and plans for mitigation? Did the results support any hypotheses made? If not, why?

Resources

- "Modern" Validity Theory
  - Cook, David A., and Thomas J. Beckman. "Current concepts in validity and reliability for psychometric instruments: theory and application." The American journal of medicine 119.2 (2006): 166-e7
  - Edwards, Michael C., et al. "Fit for purpose and modern validity theory in clinical outcomes assessment." Quality of life research 27 (2018): 1711-1720
  - Messick, S. (1989b). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan

Resources *(cntd., 1)*

- Evidential Pluralism
  - Parkkinen, VP, et. al. Evaluating Evidence of Mechanisms in Medicine: Principles and Procedures. Springer International Publishing, 2018
  - Shan, Y., Williamson, J. (2023). Evidential Pluralism in the Social Sciences. United States: Taylor & Francis

Questions & Answers

Partnership for Quality Measurement
Powered by Battelle

Risk Adjustment: Variability and Innovation

Beth Jackson | Battelle

Michelle Sunderman | Battelle

October 16, 2025
2:30PM 3:25PM (ET)

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.

Meet the Presenters

Beth Jackson | E&M Evaluation Lead

- Coordinates endorsement and maintenance (E&M) staff assessments, provides evaluation technical support
- 15+ years' research and evaluation experience

Michelle Sunderman | Data Scientist

- Leads E&M risk adjustment methods evaluation
- 4 years of quality and data science experience

Session Objectives and Agenda

- Objectives:
  - Examine current practices and challenges in risk adjustment, with a focus on variability and implementation at the entity level.
  - Explore emerging tools and techniques, including machine learning, and solicit developer feedback on opportunities for standardization.
- Agenda
  - Accounting for Risk in Clinical Quality Measures
  - Statistical Approaches and Model Performance
  - Emerging Tools and Challenges
  - Standardization and Developer Feedback
  - Q&A

Accounting for Risk in Clinical Quality Measures

Traditional Definitions

Risk adjustment and case-mix adjustment are often used interchangeably, but conceptually they differ.

Risk adjustment

Broader in scope, generally referring to adjusting for **any** patient factors that predict the outcome, regardless of their causal role.

- Typically includes both moderators and mediators unless explicitly restricted.
- In payment contexts, “risk adjustment” often refers to predicting total expected costs, not necessarily disentangling mechanisms.

Case-mix adjustment

Controls for **patient-level characteristics** present at baseline that might moderate the relationship between an intervention/exposure (e.g., provider quality) and the outcome.

- Examples: age, socioeconomic status, comorbidities, functional status.
- Goal: make entities (hospitals, clinicians) comparable by adjusting for factors *outside their control* that influence outcomes.

Mediators vs. Moderators as a Clarifying Lens

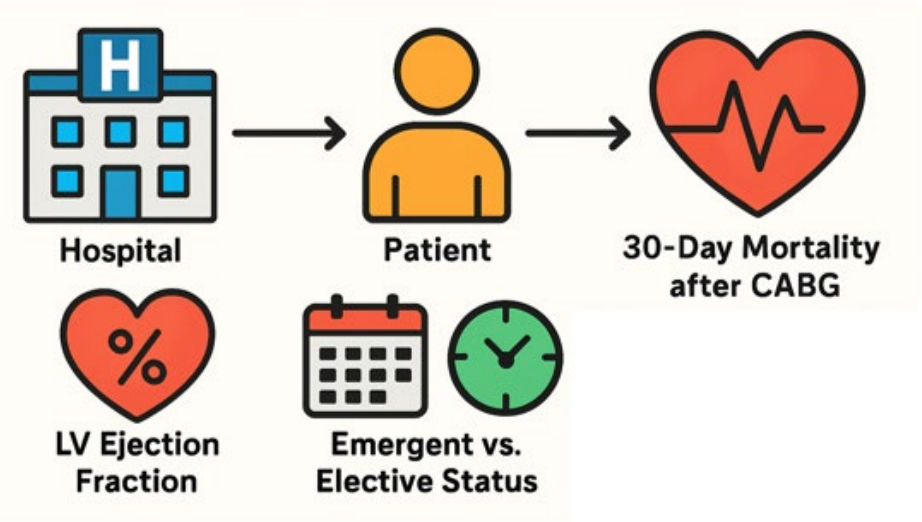
- Using mediators and moderators provides a principled way to distinguish between risk adjustment and case-mix adjustment.

Aspect	Risk Adjustment	Case-Mix Adjustment
Causal role	Adjusts for mediators (e.g., competing causes) between patient characteristics and the outcome.	Adjusts for moderators (e.g., counteracting mechanisms) that modify the effect of the provider's actions on the outcome.
Purpose	Reduce bias from alternative causal pathways unrelated to provider performance.	Reduce unfair comparisons when entities serve different populations.
Examples	Disease severity, tumor stage, pre-existing organ failure → directly drive the outcome, irrespective of provider quality.	Social support, health literacy, access to transportation → modify the effectiveness of provider quality on the outcome.
Effect on interpretation	"Given the same baseline risk, do better-performing hospitals achieve better outcomes?"	"After accounting for population differences, are outcomes comparable?"

Practical Implications for Outcome Measures

When to Prefer Risk Adjustment

- For outcome measures where **differences in baseline severity** or competing risks dominate.
- Example: 30-day mortality after Coronary Artery Bypass Graft (CABG) surgery
  - Mediators: left ventricular (LV) ejection fraction, emergent vs. elective status → directly impact mortality
  - Adjusting for these is necessary to avoid penalizing hospitals treating sicker patients.



Practical Implications for Outcome Measures (*cntd., 1*)

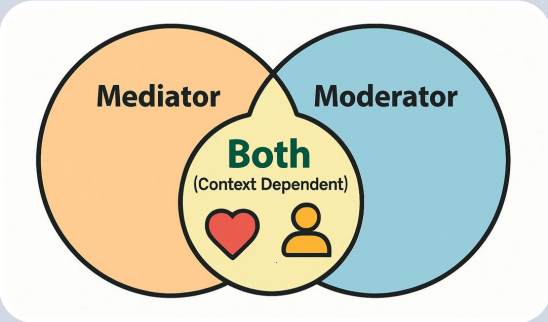
When to Prefer Case-Mix Adjustment

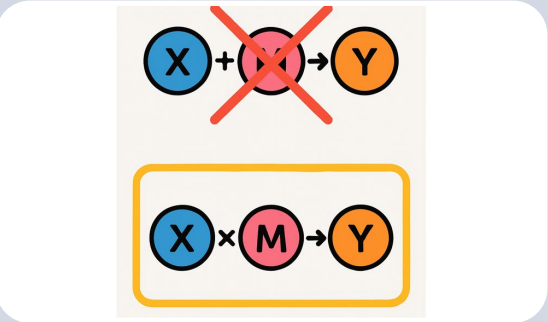
- For measures sensitive to **differential responsiveness** based on population context
- Example: Readmissions or patient-reported outcomes
  - Moderators: income, caregiver support, health literacy → determine whether provider actions translate into better outcomes
  - Adjusting here improves **fairness** but should be justified carefully to avoid adjusting away *potentially avoidable differential responses*.

"**Potentially avoidable**" invites an empirical question rather than a normative presumption:

- Which portions of the differential are proximal and modifiable by provider quality or system-level interventions? (accountability)
- Which portions are structural and persistent beyond provider influence (at least within the measurement window)? (fairness)

Caveats and Risks





Overlap is common: Many variables act as both mediator and moderator depending on the context.

- For example, diabetes can directly cause poor wound healing (mediator) *and* reduce responsiveness to postoperative education (moderator).

Policy sensitivity: Over-adjusting for moderators risks obscuring potentially avoidable differential responses, which can potentially conflict with quality program mandates.

Statistical complexity: Moderators imply interaction effects; simple additive models are insufficient.

Recommendations

- Framing can work if expressed as:
  - **Risk adjustment** → adjusts for **mediators** when variables represent **competing causes** of the outcome, not related to provider quality
  - **Case-mix adjustment** → adjusts for **moderators** when variables represent **counteracting mechanisms** that affect how patients respond to provider quality
- Make sure to emphasize:
  - This is a **causal interpretation**, not a universally accepted definition.
  - Methodologies should be aligned with DAG-based causal diagrams to make the mediator/moderator distinction explicit.

Conceptual Model

Overview and Purpose

What is a Conceptual Model?

- A causal diagram that maps out how patient clinical, functions, and contextual factors relate to the outcome
- Guides the development of risk adjustment or stratification models

Key Components

- Target population
- Measured outcome
- Risk factors and variables
- Causal pathways
- Mediators, moderators, and independent effects
- Hypothesized relationships

Sources of Input

- Literature review
- Expert opinion, including technical expert panels, workgroups, clinicians, statisticians, and methodologists

Best Practices/Tips

- Clearly map how each risk factor or variables influences the outcome
- Distinguish between mediators, moderators, and independent effects
- Prioritize risk factors based on clinical relevance and cost/benefit of data collection
- Consider all relevant risk factors, regardless of data availability, and explain potential bias from their exclusion

Conceptual Model

Selecting Data Sources and Variables

- A **risk factor** is a *concept*; a **variable** operationalizes this concept.
- Primary data source may limit available variables for one or more key risk factors.
- Additional data sources that enrich the primary data source may be needed.
- Proxy indicators and their tradeoffs should be considered carefully.
  - Potential benefits: reduced data collection cost, burden
  - Risks: ambiguous meaning, inadvertently removing variance due to quality

Risk Factor	**Variable** *(EHR)*	**Variable** *(Claims)*	**Variable** *(Survey)*
Race/ethnicity	Inconsistent, often missing	"White," "Black," "Other"	OMB categories (7)
Comorbid conditions	Drawn from the *Problem List* – can be inconsistent	ICD-10 diagnosis codes –standardized and widely used	Self-reported (recall bias)
Income	Not standardized but sometimes available (e.g., in unstructured notes); insurance (e.g., Medicaid) often used as a proxy	No direct income capture; ICD-10 Z codes for economic hardship (e.g., Z59.5 extreme poverty) but rarely used	Annual household income in categories

Statistical Approaches and Model Performance

Risk Adjustment Methods

- Modeling techniques: use statistical models (e.g., logistic regression) to estimate outcome probability based on patient characteristics
  - Flexible approach that supports continuous and categorical variables
  - Allows for covariate adjustment and hierarchical structures
- Indirect standardization: compares observed outcomes to expected outcomes, where expected is based on a model applied to a reference population
  - Accounts for case mix and useful for comparative performance
- Direct standardization: applies reference rates to the population's distribution of risk factors
  - Allows for stratified comparisons and is transparent

Common Statistical Models

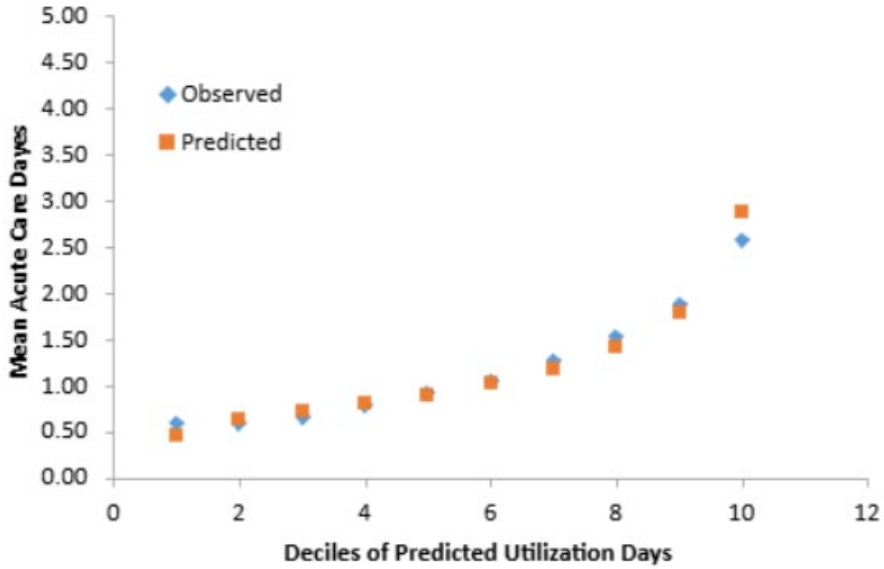
Feature	Fixed Effects Model	Random Effects Model	Mixed Effects Model
Description	Estimates a separate effect for each measured entity (e.g., hospital)	Assumes measured entity effects are randomly drawn from a distribution	Combines fixed effects with random effects
Use case example	When comparing specific measured entities in a sample	When quantifying variation across measured entities and generalizing beyond them	When adjusting for individual-level covariates and group clustering simultaneously
Generalization	Applies only to the measured entities in the dataset	Generalizes to a larger population of measured entities	Generalizes both individual and group effects
Data considerations	Requires sufficient data for each measured entity	Handles many measured entities with fewer data points per entity	Best with hierarchical/ clustered data
Strengths	Controls for all time-invariant differences; reduces bias	More efficient with a larger number of entities	Flexible; can model both within- and between-entity effects
Limitations	Cannot estimate effects of variables that do not vary within entities; less generalizable	Assumes random effects are uncorrelated with predictors	More complex to specify and interpret; requires larger sample sizes for stable estimates

Feature Selection Methods

- Stepwise regression: adds or removes variables, retaining those that are statistically significant or improve model fit
  - Limitations: overfitting, redundancy, may not consider all variable combinations
- Regularization methods (LASSO): shrinks the effects of less important variables, retaining features most related to the outcome
  - Limitations: interpretability, correlated features
- Resampling methods: combine feature selection methods with bootstrapping and retain variables most frequently selected across bootstrap datasets
- Clinical expert consultation: may inform relationships among model features and help justify decisions to retain or remove variables
  - Recommended for all risk adjustment models

Key Performance Metrics – Calibration

- Calibration: how closely predicted probabilities of outcomes align with observed outcomes
  - Calibration plots
  - Predictive ratio
  - Hosmer-Lemeshow
- High calibration means that the model is accurate in its probability estimates and patient risk is properly accounted for



Example calibration plot submitted for CBE 2881: Excess Days in Acute Care (EDAC) after Hospitalization for Acute Myocardial Infarction (AMI)

Key Performance Metrics – Discrimination

- Discrimination: how well a model can distinguish between different outcomes
  - C-statistic
  - Sensitivity
  - Specificity
- Discrimination informs how well the model can rank individuals by their risk of experiencing the outcome.
- High discrimination means the model can differentiate between “high-risk” and “low-risk” patients.

Key Performance Metrics – Goodness of Fit

- Goodness of fit: how well a model fits the entire dataset and captures variance in the data
  - Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)
    - Lower values indicate a better-fitting model
  - R-squared
    - Values range from 0 to 1, with higher values indicating the model explains more of the variability in the data

Emerging Tools and Challenges

Machine Learning

- **Machine learning** is a computational technique where models "learn" from large amounts of data to make predictions based on patterns in the information.

Supervised learning: models are trained on labeled data, meaning each input includes a known outcome. The goal is to accurately predict outcomes for new, unseen data.

- Tree-based: decision trees, random forest
- Regression-based: linear, logistic, ridge, LASSO
- Other: neural networks, support vector machines

Unsupervised learning: Models are trained using unlabeled data, without predefined outcomes. The goal is to discover hidden patterns or structures within the data.

- Clustering techniques (e.g., hierarchical clustering)
- Principal components analysis

LASSO: Least Absolute Shrinkage and Selection Operator

Machine Learning *(cntd., 1)*

Potential benefits

- Improved predictive accuracy, capture complex relationships
- Complex data handling
- Feature selection

Challenges

- Can require significant computational resources
- Sometimes difficult to interpret
- Potential for overfitting, bias

Risk Adjustment Challenges

Challenge	Mitigation Approach	Considerations
Incomplete or inaccurate data (i.e., missing or misclassified data)	Improve data collection standards; use multiple data sources (e.g., claims + EHR); validate coding practices	Consider the potential for selection bias if data completeness varies across entities; better-resourced providers may appear to perform better due to more complete data
Variable selection (i.e., difficulty distinguishing confounders, moderators, and mediators; risk of adjusting away differences)	Develop a robust conceptual model, informed from rigorous evidence and expert input; avoid adjusting for mediators	Consider any risks of omitting relevant variables or over-adjusting, which can adjust away differences or reduce model sensitivity to care effects
Unmeasured variables (i.e., lack of data capturing risk factor[s] directly)	Use proxies for unmeasured variables, such as indirect indicators (e.g., diagnosis codes for disease severity, ZIP codes for sociocontextual status)	Use granular, validated proxies; incorporate external data sources (e.g., area deprivation index, census data)

Best Practices

Develop a comprehensive conceptual model: Create a risk adjustment framework for all measures, informed by rigorous evidence (e.g., primary literature) and expert input. Clearly define mediator and moderator factors to support the model's rationale.

Ensure transparent documentation and reporting: Maintain clear and thorough documentation of risk adjustment models and data sources, including model specifications, variable definitions, and decision rationales, to facilitate transparency and reproducibility.

Support variable selection with empirical testing: Use empirical analyses to guide the inclusion or exclusion of variables in the model, ensuring that decisions are data driven and evidence based.

Utilize test and validation datasets: Develop and assess the model using separate test and validation datasets to evaluate performance and generalizability.

Address limitations with mitigation strategies: If limitations are identified, provide clear mitigation approaches and rationales, considering potential tradeoffs and their impact on model validity and fairness.

Standardization and Developer Feedback

E&M Risk Adjustment Requirements

- Rationale for adjustment: provide a clear rationale for adjusting or not adjusting for differences in patient characteristics across measured entities.
  - Required for intermediate outcome, cost/resource use, outcome, patient-reported outcome and patient-reported experience measures
- If adjusting for differences in patient characteristics
  - **Conceptual Model**
    - Includes patient characteristics that influence the measured outcome and are present at the start of care
    - Excludes factors associated with differences or inequities in care, unless a clear justification is provided for their inclusion
  - **Empirical Analysis**
    - Demonstrates variation in prevalence of patient characteristics across measured entities, contribution of these characteristics to unique variation in the outcome, impact of risk or case-mix adjustment for providers at high or low extremes of risk or case-mix, acceptable model performance

Opportunities for Standardization

- Collinearity assessment
- Description of feature selection methods with justification linked to the conceptual model
- Conceptual model components
- Sensitivity analyses for missing/unavailable variables
- Proxy variables
  - Evidence supporting their relationship to the outcome
- Model performance metrics
  - Contextualize performance by clinical condition and outcome frequency, not a fixed threshold
  - Reference benchmarks from similar models (e.g., same outcomes, related/similar measures)

References

- Centers for Medicare & Medicaid Services. (2022, May). *Risk adjustment in quality measurement* (Supplemental Material to the CMS Measures Management System Hub). https://mmshub.cms.gov/sites/default/files/Risk-Adjustment-in-Quality-Measurement.pdf
- Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. Value Health. 2015;18:137–40.
- Sherri Rose, Intersections of machine learning and epidemiological methods for health services research, *International Journal of Epidemiology*, Volume 49, Issue 6, December 2020, Pages 1763–1770, https://doi.org/10.1093/ije/dyaa035

Questions & Answers

Resources

Session Recordings

- Session recordings will be posted to the E&M Resources page by the end of October.
- The E&M staff provides technical assistance to measure developer and stewards at any time before or during the measure submission process. Contact PQMsupport@battelle.org with any questions.

Educational Materials

- The following education materials are available on the E&M Resources page:
  - Logic Model Guidance and Template
  - Closing Care Gaps Guidance
  - Reliability Guidance
  - What Good Looks Like:
    - Outcome Measure
    - Process Measure
    - Cost Measure

E&M Guidebook

- The E&M Guidebook is available for more information.

Partnership for Quality Measurement
Powered by Battelle

Thank You!

Have questions? Contact us at
PQMsupport@battelle.org