Partnership for Quality Measurement
Powered by Battelle

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

Day 1: Welcome and Introductions

Brenna Rabel | Battelle

October 15, 2025
12:00PM-12:10PM (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.

Workshop Objectives

The purpose of this 2-day virtual measure developer workshop (MDW) is to:

- Examine how measurement science and economic evaluation can inform quality improvement and accountability efforts, including strategies to address the cost and burden of measurement;
- Review methodological approaches for evaluating performance gaps, reliability, validity, and risk adjustment, and incorporate heuristics, statistical techniques, and causal reasoning to support scientifically sound measures;
- Explore challenges and best practices in the development and evaluation of cost measures, with a focus on practical implementation and alignment with quality objectives; and
- Foster peer learning by sharing key insights, challenges, and reflections, promoting discussion to support continuous improvement in measure development and evaluation.

Agenda

Day 1 – October 15, 12:00-3:15 PM

- 12:00 PM: Welcome
- 12:10 PM: Cost and Burden of Quality Measurement
- 12:40 PM: Performance Gap: Estimating Impact with Heuristics
- 1:55 PM: Break
- 2:10 PM: Presenting and Evaluating Cost Measures
- 3:10 PM: Closing

All times are listed in Eastern Time (ET).

Battelle MDW Team

- Nicole Brennan, MPH, DrPH, Executive Director
- Brenna Rabel, MPH, Deputy Director
- Jeff Geppert, JD, EdM, Measure Science Team Lead
- Quintella Bester, PMP, Senior Program Manager
- Matthew Pickering, PharmD, Endorsement and Maintenance (E&M) Task Lead
- Meridith Eastman, PhD, Pre-Rulemaking Measure Review (PRMR)/Measure Set Review (MSR) Task Lead
- Anna Michie, MHS, PMP, E&M Deputy Task Lead
- Beth Jackson, PhD, MA, E&M Evaluation Lead
- Lydia Stewart-Artz, PhD, PRMR/MSR Evaluation Lead
- Laura Aume, MS, Data Scientist IV
- Adrienne Cocci, MPH, Social Scientist III
- Stephanie Peak, PhD, Social Scientist III
- Isaac Sakyi, MSGH, Social Scientist III
- Michelle Sunderman, MS, Data Scientist II
- William White, MS, Data Scientist II
- Olivia Giles, MPH, Social Scientist II
- Elena Hughes, MS, Social Scientist II
- Sarah Rahman, Social Scientist I

Battelle MDW Support Team

Communications

- JJ Knight, MS
- Chauntel Richardson, MPH

Technical Editing

- Brittany Stojsavljevic
- Catherine McBride, MS

508 Compliance

- Lali Gentry, MPH

Graphics

- Sarah Kaukeinen

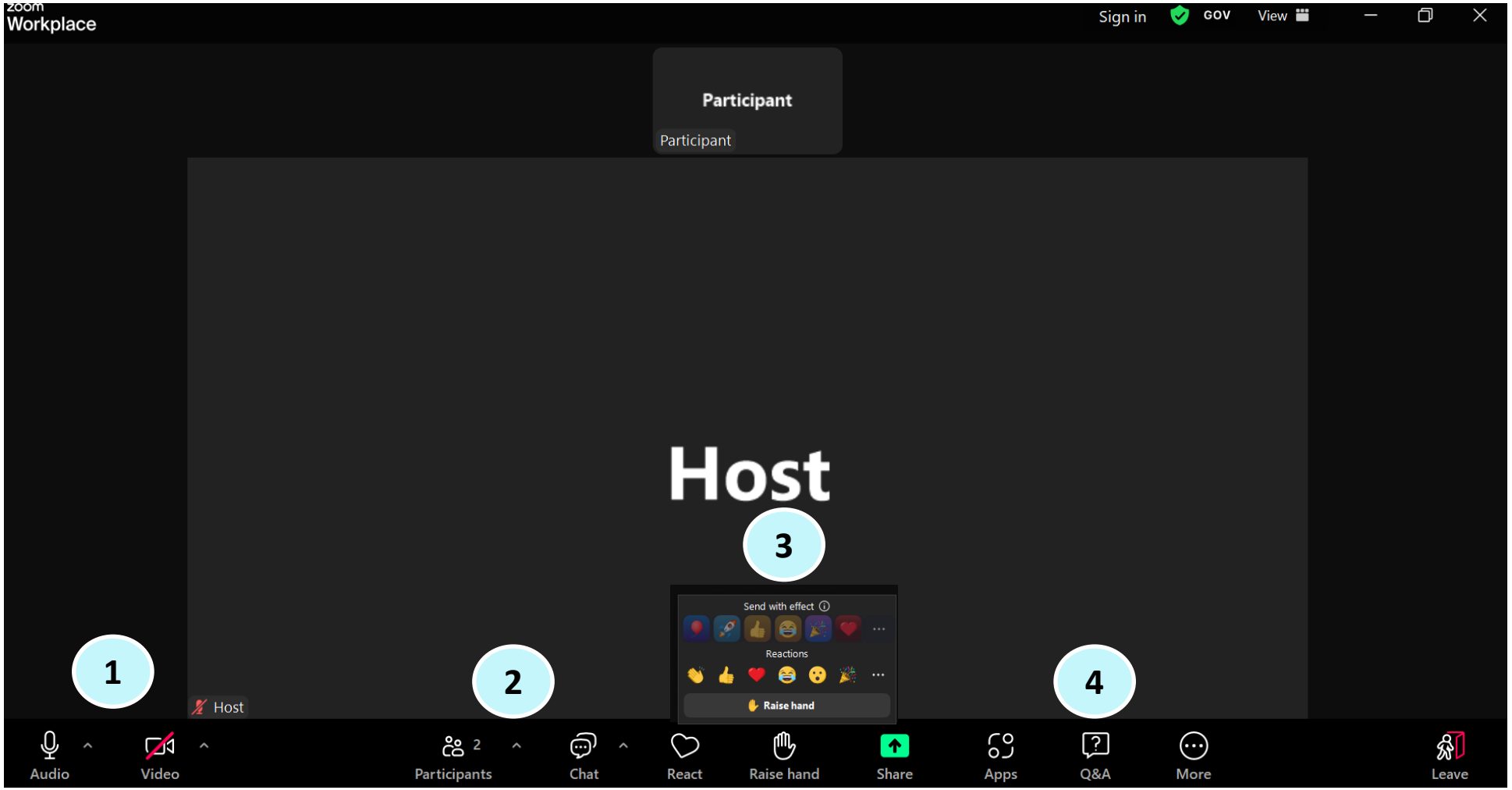
Web Team

- Kim O'Brien, MS, PMP
- Ian Warmbrodt, MBA
- Maureen Hammer, PhD, MS
- Margaret Jokoh

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



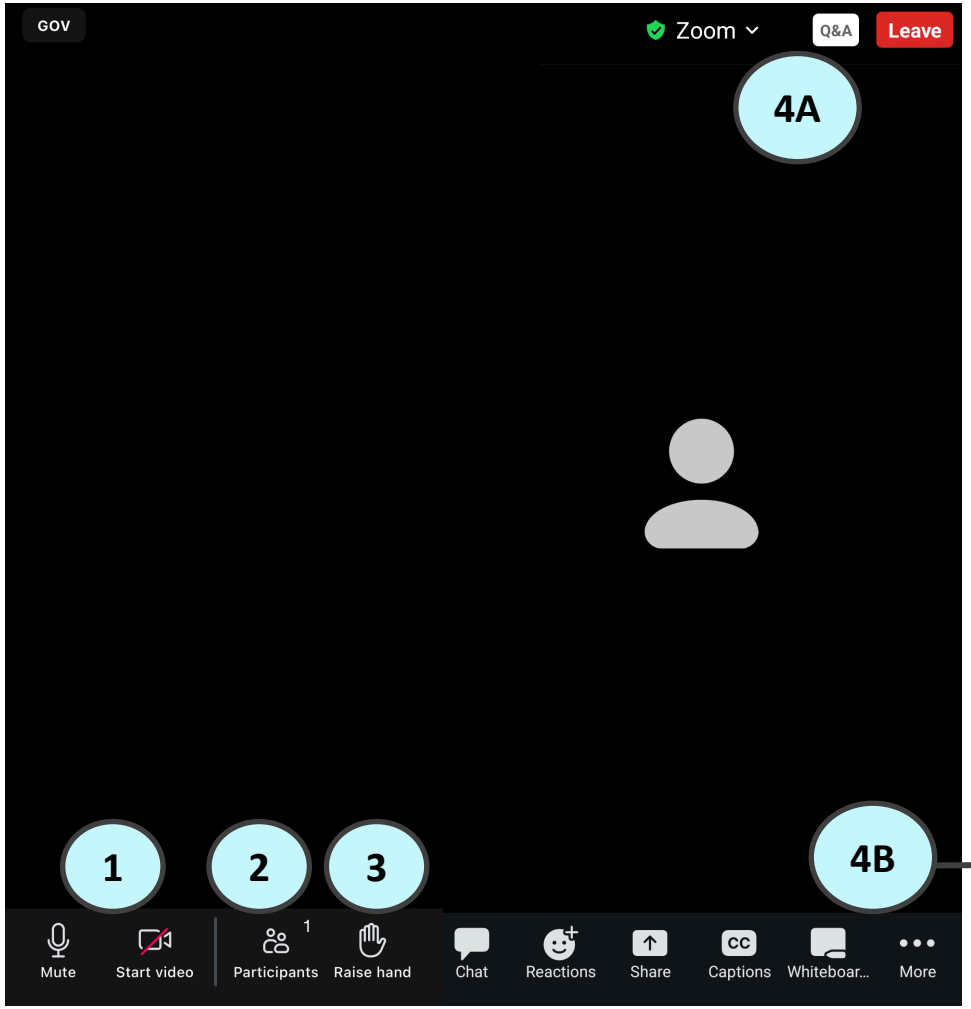
1. Click the lower part of your screen to mute/unmute or start or pause video.

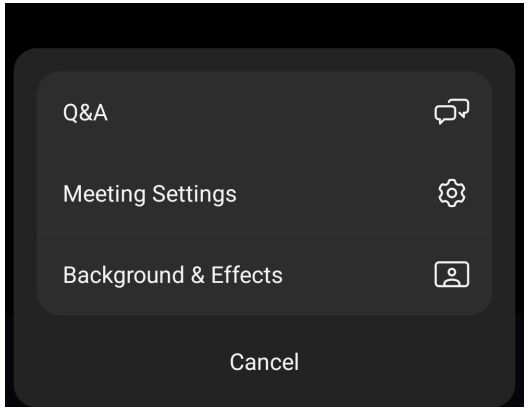
2. Click on the participant or chat button to access the full participant list or the chat box.

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).

Safety and Effectiveness

- The intention of the endorsement and maintenance (E&M) process is for measures that are safe and effective to receive endorsement.
- By “safe and effective” we mean that:
  - The measure is consistent with current professional knowledge;
  - The measure’s use is likely to improve desired health outcomes; and
  - The measure’s use is not likely to increase the risk of unintended or adverse health outcomes.
- In addition, measures should be evidence based and scientifically sound.
  - This workshop will delve deeper into aspects of scientific acceptability to support the safety and effectiveness evaluation.

E&M Enhancements

Logic Models – Building a Framework for QI Success

- Logic models visually map the relationships between resources, activities, and outcomes in health care quality improvement
  - They support the development, implementation, and assessment of both clinical quality and cost/resource use measures by outlining steps an accountable entity could take to improve performance and outcomes.
- Earlier this year, PQM developed and shared Logic Model Guidance. This guidance:
  - Includes a logic model template as a resource
  - Has led to increased consistency and more robust information in the Importance section of measure submissions

Inputs (Resources: Means)	**Activities** (What the program does: Ways)	**Outputs** (Direct results of activities)	**Outcomes** (Short-, intermediate-, and long-term)	**Impact** (Systemic changes influenced by the quality program)
			Includes the measure focus	
Feedback Mechanisms (How continuous improvement is achieved)				
Assumptions (Underlying beliefs about the quality program and context)				
External Factors (Conditions outside the quality program's control)				

Performance Gap and Reliability Decile Tables – Viewing the Larger Picture

- **Table 1: Performance Scores by Decile**

	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Mean Performance Score													
N of Entities													
N of Persons/Encounters/ Episodes													

- **Table 2: Accountable Entity-Level Reliability Testing Results by Denominator, Target Population Size**

	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability													
Mean Performance Score													
N of Entities													
N of Persons/Encounters/ Episodes													

Reliability and Validity Approaches – Strengthening the Inference

- E&M Guidebook (version 3.1) includes enhanced guidance on reliability and validity.

What's New	Why It Matters
Adds CBE recommendations for entity-level reliability methods by data type and establishes percentages for the number of entities needed to meet reliability thresholds	Promotes methodological consistency and transparency; helps developers select appropriate reliability methods and interpret results effectively
Introduces practical strategies to address and evaluate low reliability measures using strong reasoning and context	Encourages informed, risk-based decisions and enables developers to strengthen measures while minimizing potential harm
Expands guidance on entity-level validity by combining correlation analyses, causal explanation, and mechanistic evidence to support or explain observed results	Supports a more robust and theory-driven approach to validating measures, especially when empirical relationships are weaker than expected

Scientific Methods Panel – Fostering Advancements in Measurement Science

Description and Purpose

- The Scientific Methods Panel (SMP) consists of up to 15 individuals with expertise in statistics, risk adjustment, measure testing, psychometrics, economics, composite measures, and electronic clinical quality measures.
- They convene at least twice annually via virtual meetings to advise on methodologic challenges and solutions related to scientific acceptability (i.e., reliability and validity), closing gaps in care, risk adjustment, and emerging measurement approaches.
- PQM is committed to transparency: SMP members will be seated through a formal nominations process annually with a public-facing roster and meeting proceedings.

Members

- J. Matt Austin, PhD
- Daniel Deutscher, PT, PhD
- Marybeth Farquhar, PhD, MSN, RN
- Laurent Glance, MD
- Sherrie Kaplan, PhD, MPH
- Paul Kurlansky, MD
- Zhenqiu Lin, PhD
- Jack Needleman, PhD
- Sean O'Brien, PhD
- Jennifer Perloff, PhD
- Patrick Romano, MD, PhD
- Sam Simon, PhD
- Alex Sox-Harris, PhD, MS
- Ronald Walters, MD, MBA, MHA, MS
- Susan White, PhD, RHIA, CHDA

Partnership for Quality Measurement
Powered by Battelle

Cost and Burden of Quality Measurement

Donald E. Casey Jr.,
MD, MPH, MBA, MACP, FAHA, DFACMQ, DFAAPL, CPE

October 15, 2025
12:10PM-12:40PM (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 Presenter

Donald Casey Jr., MD, MPH, MBA, MACP, FAHA, DFACMQ, DFAAPL, CPE

Dr. Casey is a distinguished academic and clinician specializing in health care quality, safety, and population health. He serves as an adjunct professor at Thomas Jefferson University's College of Population Health and as an associate professor of internal medicine at Rush Medical College. Dr. Casey is also affiliate faculty at the University of Minnesota's Institute for Healthcare Informatics and a member of the faculty in the Artificial Intelligence in Cardiology Program (ATRIA). In addition to his academic roles, he contributes to the advancement of medical quality as a senior associate editor for the American Journal of Medical Quality.

Session Objective

- To examine how measurement science and economic evaluation can be used to identify which quality improvement efforts offer the greatest return on investment and explore the practical challenges of scaling and sustaining interventions.

JAMA Network Open.

Invited Commentary | Cardiology

What's Missing From Value-Based Care—Accounting for the Costs and Benefits of Quality Improvement

Donald E. Casey Jr, MD, MPH, MBA

Related article

Author affiliations and article information are listed at the end of this article.

Current estimated direct and indirect economic cost of cardiovascular diseases from 2019 to 2020 was $422.3 billion with coronary heart disease (CHD) accounting for more than 40% of deaths attributable to diseases of the heart in the US, a rate which has recently been rising again.[1] In 2023, US clinical registries noted a total of 4.3 million catheter-based interventions for CHD,[2] with a subset of 434 230 estimated for percutaneous coronary interventions performed in the US on an emergency, urgent, and elective basis.[3]

Iodinated contrast material (ICM) is administered intravascularly daily to many inpatients and outpatients around the world to increase the diagnostic accuracy of a large variety of imaging evaluations. For procedures such as cardiac catheterization procedures interventions for CHD, this type of contrast is administered intra-arterially. The most important potential major adverse effects from ICM include potentially life-threatening acute kidney injury (AKI) and allergic reactions. Many studies have shown that patients experiencing contrast induced-AKI (CI-AKI) have greater risk for death and prolonged hospitalization and other adverse outcomes, including cardiovascular events.[3]

In response, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) funded IMPROVE AKI, which was a 2 × 2 factorial cluster-randomized trial across 20 of 70 total cardiac catheterization laboratories within Veterans Affairs medical centers (VAMCs).[4] The intervention phase of the trial involved patients with preexisting chronic kidney disease, a high risk factor for both reversible and irreversible ICM-induced CI-AKI. Of note, 90% of recruited participants were male. Participating VAMCs were randomized to receive 1 of 3 complex quality improvement (QI) support strategies to assist in ensuring consistent implementation of evidence-based guideline AKI prevention recommendations through: (1) assistance; (2) assistance with either surveillance or collaborative; or (3) collaborative with surveillance. Among 122 803 patients, 13 047 experienced CI-AKI with AKI incidences of 13.3% in assistance, 11.4% in assistance with surveillance, 12.7% in collaborative, and 7.9% in collaborative with surveillance.[4]

In a recently published follow-on study, a subset of IMPROVE AKI investigators performed a well-designed, detailed cost-effectiveness analysis of these 4 implementation interventions deployed in this trial.[5] Per-patient intervention costs were (1) $12.74 for collaborative with surveillance; (2) $3.97 for collaborative alone; (3) $3.36 for assistance with surveillance; and (4) $2.69 for assistance alone. These overall results were mainly driven by costs of CI-AKI and subsequent development of permanent renal disease. Estimation of incremental cost-effectiveness ratios using willingness-to-pay threshold of $100 000 per quality-adjusted life years (QALY) revealed the collaborative with surveillance strategy as economically dominant. The authors noted overall per-person savings of $742.75 while improving effectiveness by 0.02 QALYs with this strategy.

A common response to impactful findings such as these impressive clinical results and economic evaluation of potential cost savings demonstrated by IMPROVE AKI, often goes something like this: "Why aren't we doing this in all cardiac catheterization labs everywhere?" While an excellent question, the most meaningful answer is, by no means, a simple one.

First, the US Veterans' Affairs Health Systems (VAHS) have a well-oiled and funded organization-wide structure and approach to continuous health care QI through its Quality Enhancement Research Initiative, which was first established 25 years ago and supports "evidence to action" research across the VAHS System (such as the IMPROVE AKI initiative).[6] Second, IMPROVE

Open Access. This is an open access article distributed under the terms of the CC-BY License.

JAMA Network Open. 2025;8(4):e252510. doi:10.1001/jamanetworkopen.2025.2510 April 2, 2025 1/3

Downloaded from jamanetwork.com by guest on 08/08/2025

What's Missing From Value-Based Care – Accounting for the Costs and Benefits of Quality Improvement

IMPROVE-AKI Trial

Does a Team-Based Coaching Intervention Improve Contrast-Associated AKI?

Methods

20 Veteran Affairs Medical Centers

N = 4,517 patients, 510 experienced AKI

Implementation strategies

- Technical Assistance
- Virtual Learning Collaborative
- Automated Surveillance Reporting

Randomization

2x2 factorial
Cluster randomized trial

- **Assistance**
  - 5 centers with surveillance
  - 5 centers without surveillance
- **Collaborative**
  - 5 centers with surveillance
  - 5 centers without surveillance
- **With Surveillance**
  - 5 centers with Assistance
  - 5 centers with Collaborative
- **Without Surveillance**
  - 5 centers with Assistance
  - 5 centers with Collaborative

Results

Reduction in Odds of AKI

- **Collaborative without Surveillance** (vs Assistance without surveillance): **28%** (aOR= 0.72; 0.58-0.88)
- **Assistance with Surveillance** (vs Assistance without surveillance): **24%** (aOR=0.76; 0.62-0.93)
- **Collaborative with Surveillance** (vs Assistance without surveillance): **46%** (aOR=0.54; 0.40-0.74)

Funded by NIDDK R01DK113201

Conclusions: This implementation trial estimates that the combination of Collaborative with Surveillance reduces the odds of AKI by 46% compared to Assistance alone at Veteran Affairs Medical Centers.

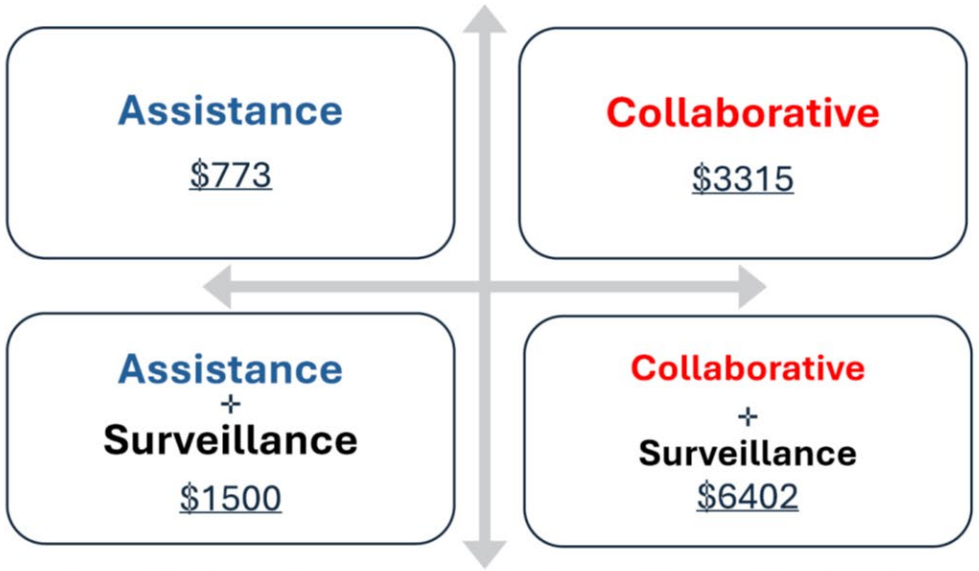
Jeremiah R. Brown, Richard Solomon, Meagan E. Stabler, et al. ***Team-Based Coaching Intervention to Improve Contrast-Associated Acute Kidney Injury.*** CJASN doi: 10.2215/CJN.0000000000000067. **Visual Abstract by Aakash Shingada, MD**

IMPROVE-AKI Trial

Economic Evaluation

Costs of Checklist Implementation Per Hospitals

18x monthly training meetings

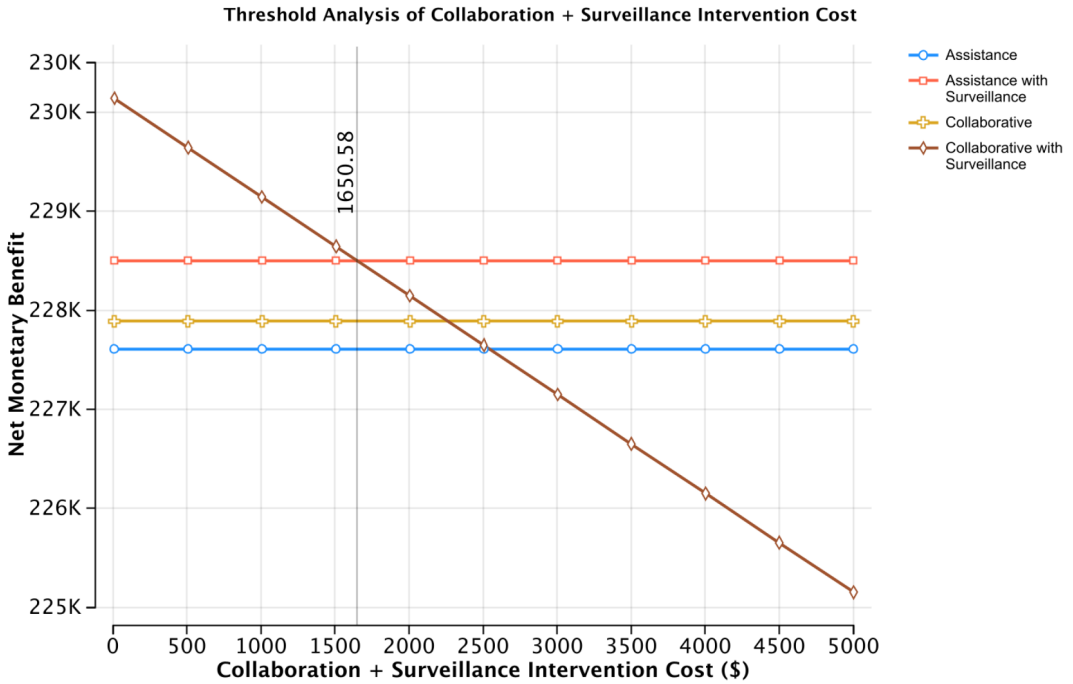


Supplement Figure S1: Costs shown represents the most expensive sites for each intervention.

IMPROVE-AKI Trial

Economic Evaluation *(cntd., 1)*

- Analysis demonstrated the greater impact of reducing AKI compared with the per-patient costs of implementing almost any strategy that improves guideline-based care.
- AKI and the subsequent development of CKD and ESRD overwhelm the wide range of realistic implementation costs associated with health care team-based interventions.
- The observed costs were attributed primarily to the cost of training, which comes out to be relatively small amount per hospital.



Supplement Figure S2: Threshold Analysis on Intervention Cost of Collaboration with Surveillance

CKI: chronic kidney disease ESRD: end stage renal disease

IMPROVE-AKI Trial
Resources

Sites

- The estimated budget for the Veterans Affairs (VA) Quality Enhancement Research Initiative (QUERI) activities in 2024 was part of the broader VA budget request of $325.1 billion.
- This included funding for various programs aimed at improving healthcare delivery and outcomes for veterans.
- QUERI focuses on implementing evidence-based practices to enhance care quality, and its funding is typically allocated within the VA's discretionary budget for research and healthcare improvement.

Trial Support

- IMPROVE AKI Trial was supported by the National Institutes of Diabetes and Digestive Kidney Diseases (NIDDK)
- Total funding: $675,270
- Translating commonly funded quality improvement research into consistent, daily practice often remains an elusive tangible outcome
- Investigators acknowledge concerns regarding sustainability for the participating catheter laboratories and generalizability work to other non-VAHS systems.

Challenges in Translating QI into Practice

- Studies like IMPROVE AKI are rare and valuable:
  - Few QI studies include rigorous economic evaluations.
    - For example, a recent scoping review of how US hospitals have objectively "bent the cost curve" through integrated QI hospital revealed that, of 4198 articles, only 19 case studies from 5 countries were identified.
  - Even fewer show clear return on investment despite well-designed interventions.
    - A study examining the economic impact of quality improvement collaboratives (QICs) found that very few published studies addressed costs or economic evaluations, despite widespread use of QICs in U.S. health care. Out of 8,505 citations, only 8 met inclusion criteria, and of those, 5 were high-quality and supported QICs as cost-effective.

Challenges in Translating QI into Practice *(cntd., 1)*

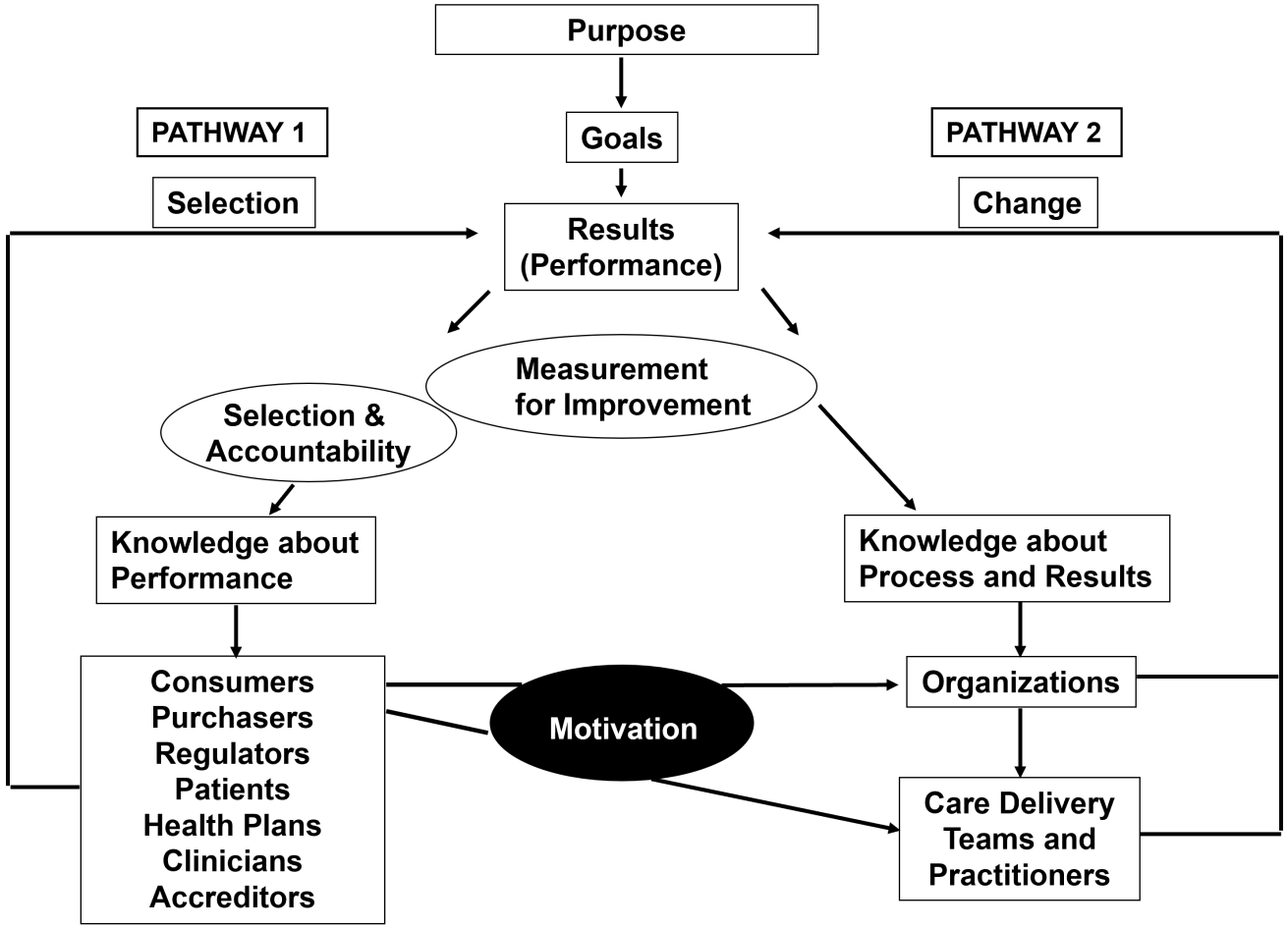
- Persistent barriers to implementation:
  - Lack of sustained physician engagement
  - Insufficient infrastructure for data collection and analysis
  - Limited resources to support full-cycle QI (design → delivery → evaluation → scale)
  - Payers and policymakers often do not incentivize or reimburse for QI-driven improvements

Challenges in Translating QI into Practice *(cntd., 2)*

Common examples of the tangible and intangible costs of real-life QI

- Salary allocations for dedicated QI personnel within quality and patient safety improvement departments, departmental and service line professionals, and unit/microsystem managers
- Other relevant personnel support from internal service departments (e.g., Six Sigma/Lean, biostatistics, healthcare analytics, electronic medical record teams, QI registries)
- External quality reporting, accreditation and certification fees
- QI research, dissemination and publication expertise and infrastructure
- Data management, cybersecurity, report generation, QI dashboard development and maintenance
- Development, implementation, evaluation and maintenance of guideline-based digital clinical decision support workflow applications
- Executive leadership management quality performance-based compensation prerequisites

Two Pathways to Quality Improvement



Opportunities For Progress

- Expand funding for QI trials that include *economic outcomes*
- Foster implementation science that supports real-world adoption of proven interventions
- Create financial signals from payers that reward sustained QI impact
- Establish standardized frameworks to evaluate and compare QI strategies across diverse settings
- **Performance and quality measurement deployment & implementation must demonstrate both health improvement impact and net economic value**

Questions & Answers

Partnership for Quality Measurement
Powered by Battelle

Performance Gap: Estimating Impact with Heuristics

Anna Michie | Battelle

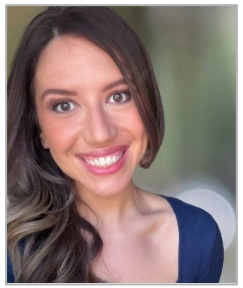
Jeffrey Geppert | Battelle

October 15, 2025
12:40PM-1:55PM (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

Anna Michie | E&M Deputy Task Lead



- Provides strategic and technical support on E&M processes and activities
- 10+ years' quality experience

Jeffrey Geppert | Sr. Research Leader

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

Session Objectives and Agenda

- Objectives:
  - Explore how heuristics and empirical methods can be used to better estimate the potential impact of quality measures
  - Examine approaches for applying causal reasoning, counterfactual modeling, and “Rule of Three" heuristics to assess performance gap
- Agenda
  - Key Concepts
  - Applying Heuristics
  - Case Studies
  - Closing Thoughts
  - Q&A

Key Concepts

Evaluating Importance (Impact)

- The claim of Importance requires that the benefit of the measure exceeds the burden of data collection, reporting, and use.
- Measure submissions rarely explicitly **quantify** benefit or burden.
- Table 1: Performance Score by Deciles (collected during Full Measure Submission) may be used to approximate benefit.

	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Mean Performance Score													
N of Entities													
N of Persons / Encounters / Episodes													

Rule of Three Heuristic

- The “Rule of Three” heuristic may be used to approximate the benefit-burden trade-off
  - Heuristic: serving to indicate or point out; stimulating interest as a means of further investigation
  - Benefit exceeds the burden if the plausible, achievable decrease (increase) in the number of adverse events (positive events) of the measure focus is at least 3X the number of entities reporting
    - Another way of expressing the rule is that on average each entity must have at least three “plausible, achievable” events
    - For example, if a measure has 3,000 entities reporting there must be at least 9,000 “plausible, achievable” events
  - Otherwise, the rebuttable presumption is that burden of reporting exceeds the benefit

Rule of Three Justification

- Justification for the “Rule of Three” heuristic is based on the limited literature on the costs of data collection, reporting, and use.
  - Saraswathula (2023) reported that in an acute inpatient hospital setting, the average cost of quality reporting per measure was $32,000. If the average cost of an avoidable utilization (e.g., readmission) is $14,000 (FY25 dollars), then that suggests three such adverse events must be prevented for the quality reporting cost to have a positive return-on-investment (ROI).
  - Richman (2022) and Tseng (2018) reported that across various settings (primary care, emergency department, ambulatory surgery, general inpatient, surgical inpatient) the average cost of administrative tasks required by quality reporting was approximately 12.8% of revenue. If approximately 25-30% of the revenue is avoidable utilization (based on estimates of health care waste by Shrank [2019]) that also suggests two or three adverse events must be prevented for the quality reporting cost to have a positive ROI.

Rule of Three Justification *(cntd., 1)*

Ruling out false, random, and structural zeros

- In statistical analysis, the "Rule of Three" states that if a certain event did not occur in a sample with n subjects, the interval from 0 to 3/n is a 95% confidence interval for the rate of occurrences in the population.
- Three is often used in measure specifications as a minimum to rule out random occurrences.

False Zeros	True Zeros
• Observer error (event occurred but was not observed) • Error in experimental design (sample different than the population; a zero value not part of the hypothesis)	• Structural zeros: associated with the system under study (a zero value is part of the hypothesis) • Random zeros: due to sampling variability, including those events that could potentially occur but did not (no mechanistic explanation)

Rule of Three Rebuttal

- The intent of the “Rule of Three” heuristic is to focus investigation on those measures where the benefit-burden trade-off is less certain.
- Various arguments might be made to rebut the presumption that burden exceeds benefit:
  - Benefit is greater. The only benefit considered by the rule is the avoidable utilization. Other benefits might be reduced risk of subsequent complications or reduced time-from-work or other benefits to the person.
  - Burden is lesser. The only burden considered by the rule is the administrative costs of quality data collection, reporting, and use. Measures that use clinical data that do not require modifications to the workflow might cost less to collect, report, and use.
  - Within the measure submission, use the measure burden narrative (Feasibility section) to capture this context.

Applying Heuristics

Importance Table

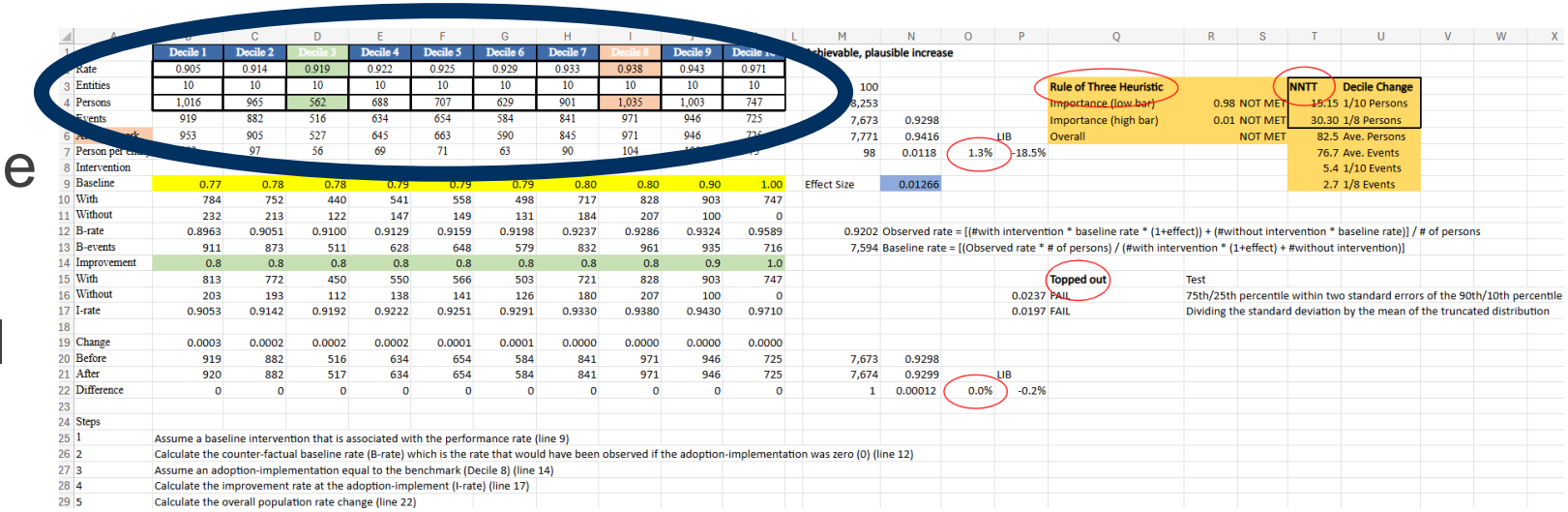
Example: CBE2023-4440e. Percent of hospitalized pneumonia patients with chest imaging confirmation (Developer: The University of Utah)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10		Achievable, plausible increase											
2	Rate	0.905	0.914	0.919	0.922	0.925	0.929	0.933	0.938	0.943	0.971													
3	Entities	10	10	10	10	10	10	10	10	10	10		100				Rule of Three Heuristic			NNTT	Decile Change			
4	Persons	1,016	965	562	688	707	629	901	1,035	1,003	747		8,253				Importance (low bar)	0.98	NOT MET	15.15	1/10 Persons			
5	Events	919	882	516	634	654	584	841	971	946	725		7,673	0.9298			Importance (high bar)	0.01	NOT MET	30.30	1/8 Persons			
6	At benchmark	953	905	527	645	663	590	845	971	946	725		7,771	0.9416		LIB	Overall		NOT MET	82.5	Ave. Persons			
7	Person per entity	102	97	56	69	71	63	90	104	100	75		98	0.0118	1.3%	-18.5%				76.7	Ave. Events			
8	Intervention																			5.4	1/10 Events			
9	Baseline	0.77	0.78	0.78	0.79	0.79	0.79	0.80	0.80	0.90	1.00		Effect Size	0.01266						2.7	1/8 Events			
10	With	784	752	440	541	558	498	717	828	903	747													
11	Without	232	213	122	147	149	131	184	207	100	0													
12	B-rate	0.8963	0.9051	0.9100	0.9129	0.9159	0.9198	0.9237	0.9286	0.9324	0.9589		0.9202	Observed rate = [(#with intervention * baseline rate * (1+effect)) + (#without intervention * baseline rate)] / # of persons										
13	B-events	911	873	511	628	648	579	832	961	935	716		7,594	Baseline rate = [(Observed rate * # of persons) / (#with intervention * (1+effect) + #without intervention)]										
14	Improvement	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	1.0													
15	With	813	772	450	550	566	503	721	828	903	747						Topped out		Test					
16	Without	203	193	112	138	141	126	180	207	100	0					0.0237	FAIL		75th/25th percentile within two standard errors of the 90th/10th percentile					
17	I-rate	0.9053	0.9142	0.9192	0.9222	0.9251	0.9291	0.9330	0.9380	0.9430	0.9710					0.0197	FAIL		Dividing the standard deviation by the mean of the truncated distribution					
18																								
19	Change	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000													
20	Before	919	882	516	634	654	584	841	971	946	725		7,673	0.9298										
21	After	920	882	517	634	654	584	841	971	946	725		7,674	0.9299		LIB								
22	Difference	0	0	0	0	0	0	0	0	0	0		1	0.00012	0.0%	-0.2%								
23																								
24	Steps																							
25	1	Assume a baseline intervention that is associated with the performance rate (line 9)																						
26	2	Calculate the counter-factual baseline rate (B-rate) which is the rate that would have been observed if the adoption-implementation was zero (0) (line 12)																						
27	3	Assume an adoption-implementation equal to the benchmark (Decile 8) (line 14)																						
28	4	Calculate the improvement rate at the adoption-implement (I-rate) (line 17)																						
29	5	Calculate the overall population rate change (line 22)																						

Importance Table – Benchmark

- Deciles sorted by performance score
- Reports the number of entities and persons per decile
- If lower is better, the benchmark is the third decile
- If higher is better, the benchmark is the eight decile
- Performance “at the benchmark” occurs when all entities perform at the benchmark or better

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Rate	0.905	0.914	0.919	0.922	0.925	0.929	0.933	0.938	0.943	0.971
Entities	10	10	10	10	10	10	10	10	10	10
Persons	1,016	965	562	688	707	629	901	1,035	1,003	747
Events	919	882	516	634	654	584	841	971	946	725
At benchmark	953	905	527	645	663	590	845	971	946	725
Person per entity	102	97	56	69	71	63	90	104	100	75



	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Rate	0.905	0.914	0.919	0.922	0.925	0.929	0.933	0.938	0.943	0.971
Entities	10	10	10	10	10	10	10	10	10	10
Persons	1,016	965	562	688	707	629	901	1,035	1,003	747
Events	919	882	516	634	654	584	841	971	946	725
At benchmark	953	905	527	645	663	590	845	971	946	725
Person per entity	102	97	56	69	71	63	90	104	100	75
Intervention										
Baseline	0.77	0.78	0.78	0.79	0.79	0.79	0.80	0.80	0.90	1.00
With	784	752	440	541	558	498	717	828	903	747
Without	232	213	122	147	149	131	184	207	100	0
B-rate	0.8963	0.9051	0.9100	0.9129	0.9159	0.9198	0.9237	0.9286	0.9324	0.9589
B-events	911	873	511	628	648	579	832	961	935	716
Improvement	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	1.0
With	813	772	450	550	566	503	721	828	903	747
Without	203	193	112	138	141	126	180	207	100	0
I-rate	0.9053	0.9142	0.9192	0.9222	0.9251	0.9291	0.9330	0.9380	0.9430	0.9710
Change	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000
Before	919	882	516	634	654	584	841	971	946	725
After	920	882	517	634	654	584	841	971	946	725
Difference	0	0	0	0	0	0	0	0	0	0

achievable, plausible increase

100
8,253
7,673 0.9298
7,771 0.9416 LIB
98 0.0118 1.3% -18.5%

Effect Size 0.01266

0.9202 Observed rate = [(#with intervention * baseline rate * (1+effect)) + (#without intervention * baseline rate)] / # of persons
7,594 Baseline rate = [(Observed rate * # of persons) / (#with intervention * (1+effect) + #without intervention)]

Rule of Three Heuristic			NNTT	Decile Change
Importance (low bar)	0.98	NOT MET	15.15	1/10 Persons
Importance (high bar)	0.01	NOT MET	30.30	1/8 Persons
Overall		NOT MET	82.5	Ave. Persons
			76.7	Ave. Events
			5.4	1/10 Events
			2.7	1/8 Events

Topped out

Test	
0.0237	FAIL
0.0197	FAIL

75th/25th percentile within two standard errors of the 90th/10th percentile
Dividing the standard deviation by the mean of the truncated distribution

7,673 0.9298
7,674 0.9299 LIB
1 0.00012 0.0% -0.2%

Steps	
1	Assume a baseline intervention that is associated with the performance rate (line 9)
2	Calculate the counter-factual baseline rate (B-rate) which is the rate that would have been observed if the adoption-implementation was zero (0) (line 12)
3	Assume an adoption-implementation equal to the benchmark (Decile 8) (line 14)
4	Calculate the improvement rate at the adoption-implement (I-rate) (line 17)
5	Calculate the overall population rate change (line 22)

Importance Table – Number Needed to Treat

- Performance at the benchmark is an increase of 98 events (1.3%)
- Rule of Three heuristic is 0.98 (98 events per 100 entities) which is less than 3
- Number needed to treat (NNTT) is 15.15 if the intervention effectiveness is calculated from the first to tenth decile
- NNTT is 30.30 if effectiveness is calculated from first to eighth decile (benchmark)

M	N	O	P	Q	R	S	T	U
Achievable, plausible increase								
							NNTT	**Decile Change**
100				**Rule of Three Heuristic**				
8,253				Importance (low bar)	0.98	NOT MET	15.15	1/10 Persons
7,673	0.9298			Importance (high bar)	0.01	NOT MET	30.30	1/8 Persons
7,771	0.9416		LIB	Overall		NOT MET	82.5	Ave. Persons
98	0.0118	1.3%	-18.5%				76.7	Ave. Events
							5.4	1/10 Events
Effect Size	0.01266						2.7	1/8 Events

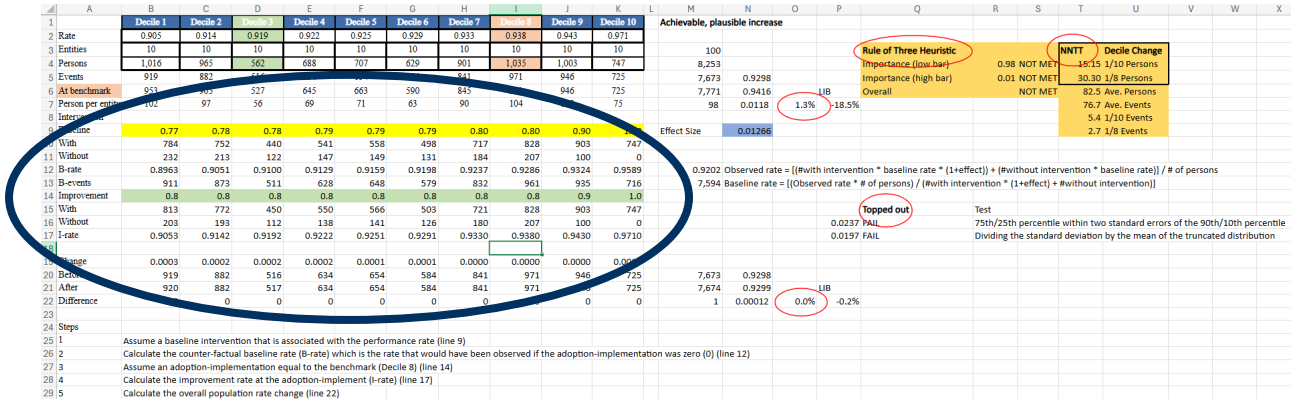
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10		Achievable, plausible increase											
2	Rate	0.905	0.914	0.919	0.922	0.925	0.929	0.933	0.938	0.943	0.971													
3	Entities	10	10	10	10	10	10	10	10	10	10		100				Rule of Three Heuristic			NNTT	Decile Change			
4	Persons	1,016	965	562	688	707	629	901	1,035	1,003	747		8,253				Importance (low bar)	0.98	NOT MET	15.15	1/10 Persons			
5	Events	919	882	516	634	654	584	841	971	946	725		7,673	0.9298			Importance (high bar)	0.01	NOT MET	30.30	1/8 Persons			
6	At benchmark	953	905	527	645	663	590	845	971	946	725		7,771	0.9416		LIB	Overall		NOT MET	82.5	Ave. Persons			
7	Person per entity	102	97	56	69	71	63	90	104	100	75		98	0.0118	1.3%	-18.5%				76.7	Ave. Events			
8	Intervention																			5.4	1/10 Events			
9	Baseline	0.77	0.78	0.78	0.79	0.79	0.79	0.80	0.80	0.90	1.00		Effect Size	0.01266						2.7	1/8 Events			
10	With	784	752	440	541	558	498	717	828	903	747													
11	Without	232	213	122	147	149	131	184	207	100	0													
12	B-rate	0.8963	0.9051	0.9100	0.9129	0.9159	0.9198	0.9237	0.9286	0.9324	0.9589		0.9202	Observed rate = [(#with intervention * baseline rate * (1+effect)) + (#without intervention * baseline rate)] / # of persons										
13	B-events	911	873	511	628	648	579	832	961	935	716		7,594	Baseline rate = [(Observed rate * # of persons) / (#with intervention * (1+effect) + #without intervention)]										
14	Improvement	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	1.0													
15	With	813	772	450	550	566	503	721	828	903	747							Topped out		Test				
16	Without	203	193	112	138	141	126	180	207	100	0						0.0237	FAIL		75th/25th percentile within two standard errors of the 90th/10th percentile				
17	I-rate	0.9053	0.9142	0.9192	0.9222	0.9251	0.9291	0.9330	0.9380	0.9430	0.9710						0.0197	FAIL		Dividing the standard deviation by the mean of the truncated distribution				
18																								
19	Change	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000													
20	Before	919	882	516	634	654	584	841	971	946	725		7,673	0.9298										
21	After	920	882	517	634	654	584	841	971	946	725		7,674	0.9299		LIB								
22	Difference	0	0	0	0	0	0	0	0	0	0		1	0.00012	0.0%	-0.2%								
23																								
24	Steps																							
25	1	Assume a baseline intervention that is associated with the performance rate (line 9)																						
26	2	Calculate the counter-factual baseline rate (B-rate) which is the rate that would have been observed if the adoption-implementation was zero (0) (line 12)																						
27	3	Assume an adoption-implementation equal to the benchmark (Decile 8) (line 14)																						
28	4	Calculate the improvement rate at the adoption-implement (I-rate) (line 17)																						
29	5	Calculate the overall population rate change (line 22)																						

Importance Table – Counterfactual

"Backs out" the adoption, implementation, and effectiveness:

1. Assume a baseline adoption-implementation associated with performance rate (9)
2. Calculate the "reference" counterfactual rate (B-rate) if the adoption-implementation was zero and calculated effectiveness (12)
3. Assume an adoption-implementation equal to the benchmark (Decile 8) (14)
4. Calculate the "intervention" counterfactual rate at the benchmark adoption-implementation (I-rate)(17)
5. Calculate the population rate change (22)

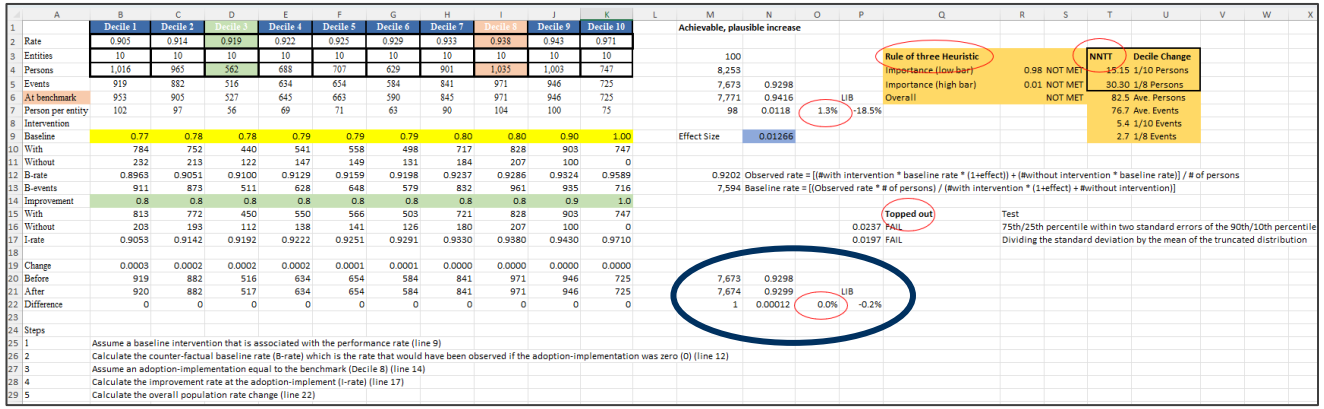
8	Intervention										
9	Baseline	0.77	0.78	0.78	0.79	0.79	0.79	0.80	0.80	0.90	1.00
10	With	784	752	440	541	558	498	717	828	903	747
11	Without	232	213	122	147	149	131	184	207	100	0
12	B-rate	0.8963	0.9051	0.9100	0.9129	0.9159	0.9198	0.9237	0.9286	0.9324	0.9589
13	B-events	911	873	511	628	648	579	832	961	935	716
14	Improvement	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	1.0
15	With	813	772	450	550	566	503	721	828	903	747
16	Without	203	193	112	138	141	126	180	207	100	0
17	I-rate	0.9053	0.9142	0.9192	0.9222	0.9251	0.9291	0.9330	0.9380	0.9430	0.9710
18											
19	Change	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000
20	Before	919	882	516	634	654	584	841	971	946	725
21	After	920	882	517	634	654	584	841	971	946	725
22	Difference	0	0	0	0	0	0	0	0	0	0



Importance Table – Achievable, Plausible

- Counterfactual suggests that little improvement is achievable and plausible
- Burden of measurement likely exceeds benefit (Rule of Three equal to 0.01)
- Rebuttable presumption of adoption, implementation, or effectiveness from literature or experience
- Re-focus or change the target population with a greater benefit

	M	N	O	P	Q	R	S	T	U
	Achievable, plausible increase								
								NNTT	**Decile Change**
	100				Rule of three Heuristic			15.15	1/10 Persons
	8,253				Importance (low bar)	0.98	NOT MET	30.30	1/8 Persons
	7,673	0.9298			Importance (high bar)	0.01	NOT MET	82.5	Ave. Persons
	7,771	0.9416		LIB	Overall		NOT MET	76.7	Ave. Events
	98	0.0118	1.3%	-18.5%				5.4	1/10 Events
								2.7	1/8 Events
	Effect Size	0.01266							



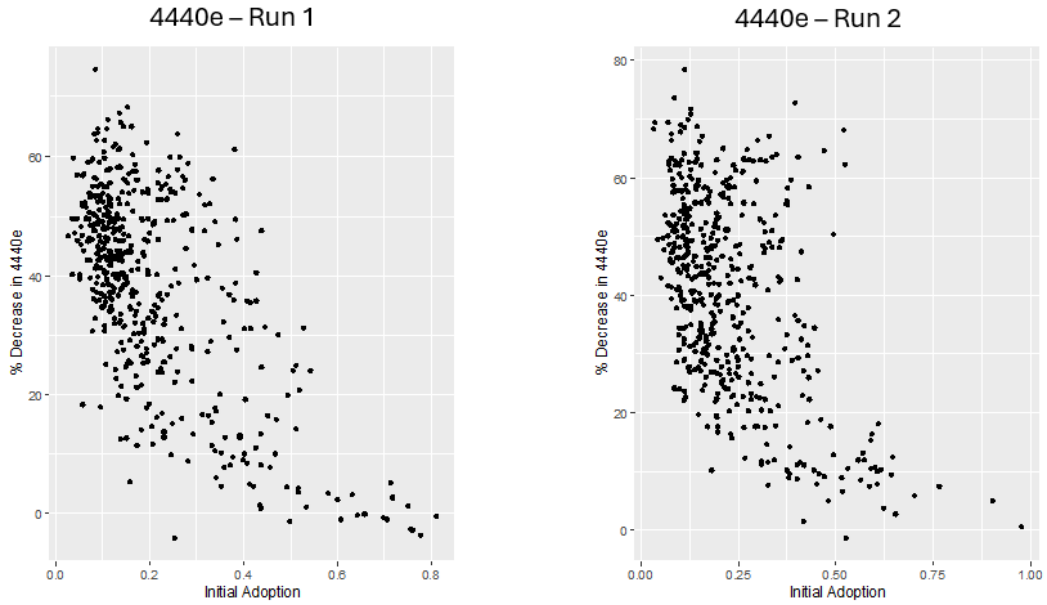
	A	B	C	D	E	F	G	H	I	J	K
1		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
2	Rate	0.905	0.914	0.919	0.922	0.925	0.929	0.933	0.938	0.943	0.971
3	Entities	10	10	10	10	10	10	10	10	10	10
4	Persons	1,016	965	562	688	707	629	901	1,035	1,003	747
5	Events	919	882	516	634	654	584	841	971	946	725
6	At benchmark	953	905	527	645	663	590	845	971	946	725
7	Person per entity	102	97	56	69	71	63	90	104	100	75
8	Intervention										
9	Baseline	0.77	0.78	0.78	0.79	0.79	0.79	0.80	0.80	0.90	1.00
10	With	784	752	440	541	558	498	717	828	903	747
11	Without	232	213	122	147	149	131	184	207	100	0
12	B-rate	0.8963	0.9051	0.9100	0.9129	0.9159	0.9198	0.9237	0.9286	0.9324	0.9589
13	B-events	911	873	511	628	648	579	832	961	935	716
14	Improvement	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	1.0
15	With	813	772	450	550	566	503	721	828	903	747
16	Without	203	193	112	138	141	126	180	207	100	0
17	I-rate	0.9053	0.9142	0.9192	0.9222	0.9251	0.9291	0.9330	0.9380	0.9430	0.9710
18											
19	Change	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000
20	Before	919	882	516	634	654	584	841	971	946	725
21	After	920	882	517	634	654	584	841	971	946	725
22	Difference	0	0	0	0	0	0	0	0	0	0
23											
24	Steps										
25	1	Assume a baseline intervention that is associated with the performance rate (line 9)									
26	2	Calculate the counter-factual baseline rate (B-rate) which is the rate that would have been observed if the adoption-implementation was zero (0) (line 12)									
27	3	Assume an adoption-implementation equal to the benchmark (Decile 8) (line 14)									
28	4	Calculate the improvement rate at the adoption-implement (i-rate) (line 17)									
29	5	Calculate the overall population rate change (line 22)									

Importance Table – Simulation

- Simulation considers the range of values of adoption, implementation, and effectiveness given the performance rate, number of entities, and number of persons
- Separate simulations for adoption, implementation, effectiveness, and overall performance
  - Adoption: % of entities
  - Implementation: % of persons within entities
  - Effectiveness: %decrease or %increase (odds ratio)

Test: Decrease in metric as initial adoption rate increases.

Reasoning: Because initial adoption is higher, there is less increase to 80% overall implementation. If the system is already close to the goal, then there is less improvement.

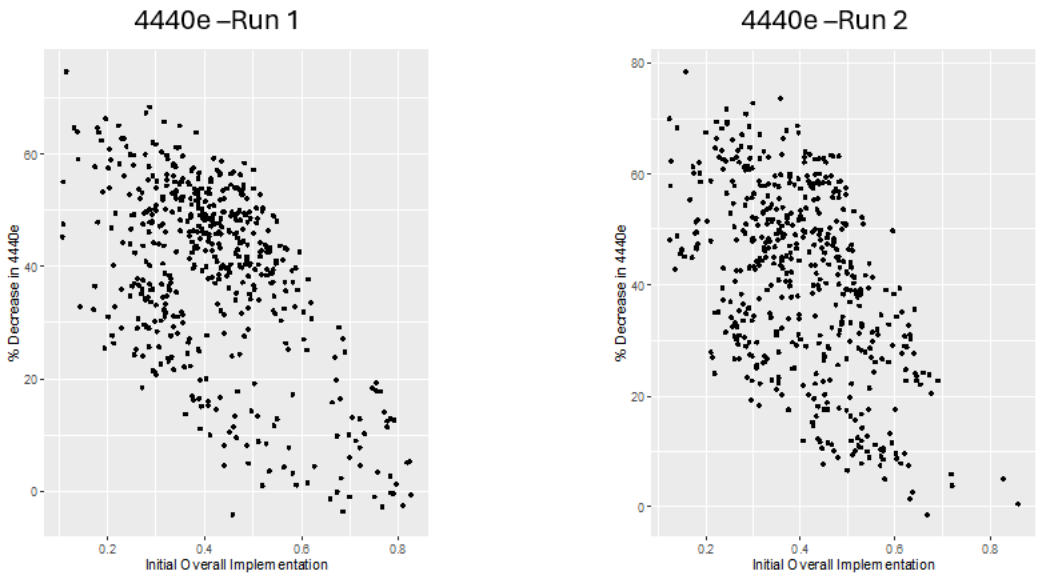


Adoption

Importance Table – Simulation *(cntd., 1)*

Test: Decrease in metric as initial overall mean implementation rate increases, where implementation without adoption is also included.

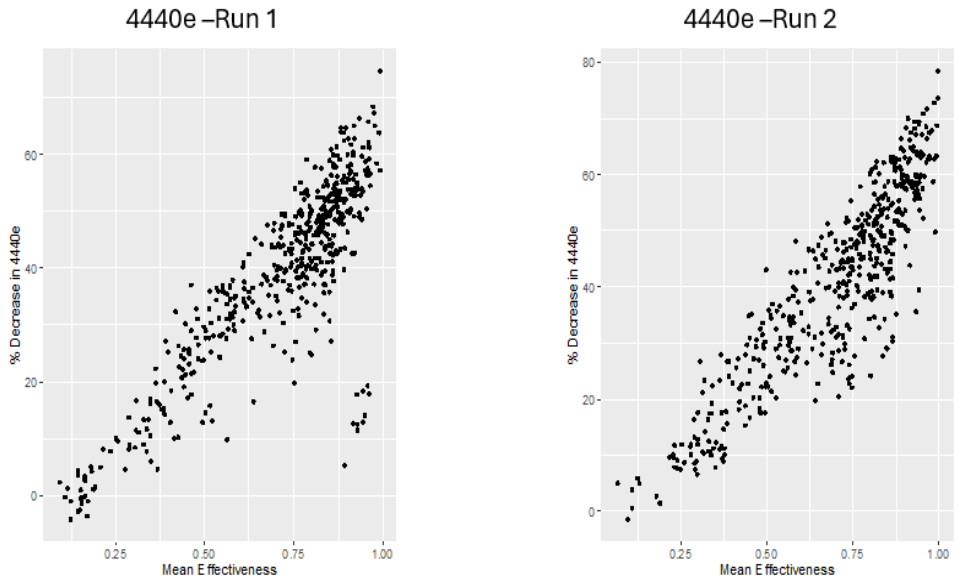
Reasoning: Because initial mean overall implementation is higher, there is even less increase to 80% overall implementation since this also considers implementation in non-adopting units. If the system is already close to the goal, then there is less improvement.



Implementation

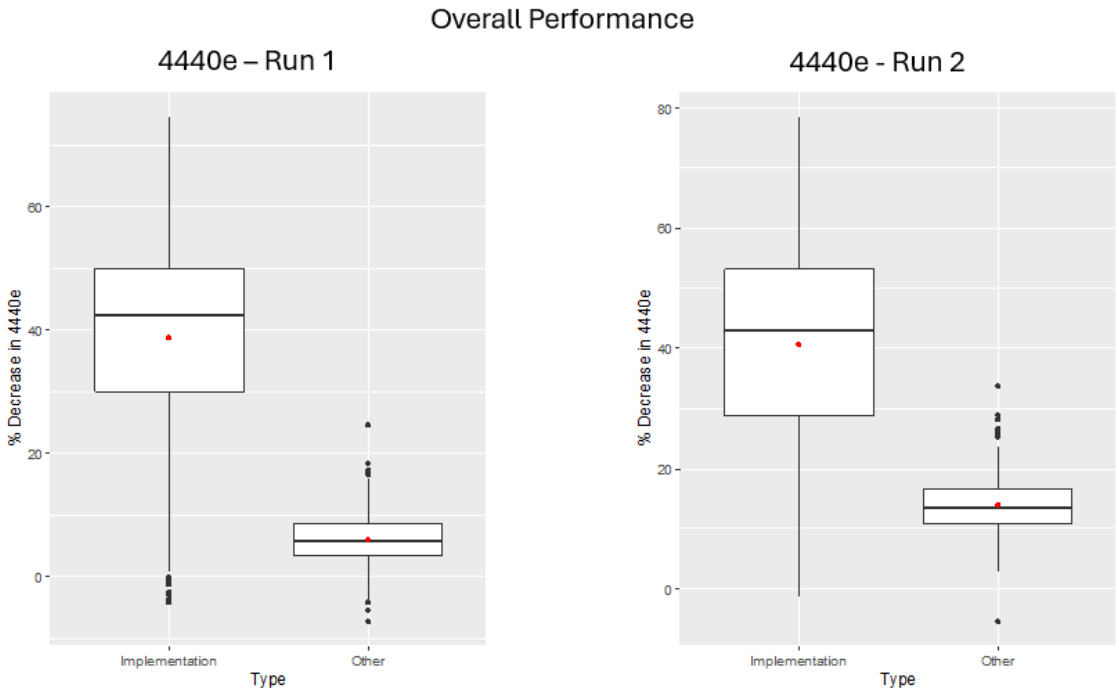
Test: Increase in metric as effectiveness increases.

Reasoning: If the intervention is less effective then there will be less benefit to the intervention.



Effectiveness

Importance Table – Simulation *(cntd., 2)*



Overall Performance

Simulation	LIB Benchmark	LB	UB	LIB Counterfactual	LB	UB
percent	0.43000	0.30000	0.50000	0.07500	0.05000	0.10000
exp()	0.65051	0.74082	0.60653	0.92774	0.95123	0.90484
base rate	0.07020	0.07020	0.07020	0.07020	0.07020	0.07020
new rate	0.04567	0.05201	0.04258	0.06513	0.06678	0.06352
LIB	-43.0%	-30.0%	-50.0%	-7.5%	-5.0%	-10.0%
base rate	0.92980	0.92980	0.92980	0.92980	0.92980	0.92980
new rate	0.95433	0.94799	0.95742	0.93487	0.93322	0.93648
HIB	2.6%	1.9%	2.9%	0.5%	0.4%	0.7%

- Simulation results for higher is better (HIB) must be transformed from lower is better (LIB)
- Performance at the benchmark (eighth decile) would be an increase of 2.6%
- Achievable, plausible performance would be an increase of 0.5%

Questions & Answers

Case Studies

Case Studies

How does the counterfactual estimate correspond to actual performance and published literature?

- Screening for Metabolic Disorders
- 30-Day Unplanned Hospital Readmission
- Influenza Immunization
- Diabetes Mellitus: Diabetic Foot and Ankle Care, Peripheral Neuropathy – Neurological Evaluation (process measure)
- Death Rate Among Surgical Inpatients with Complications

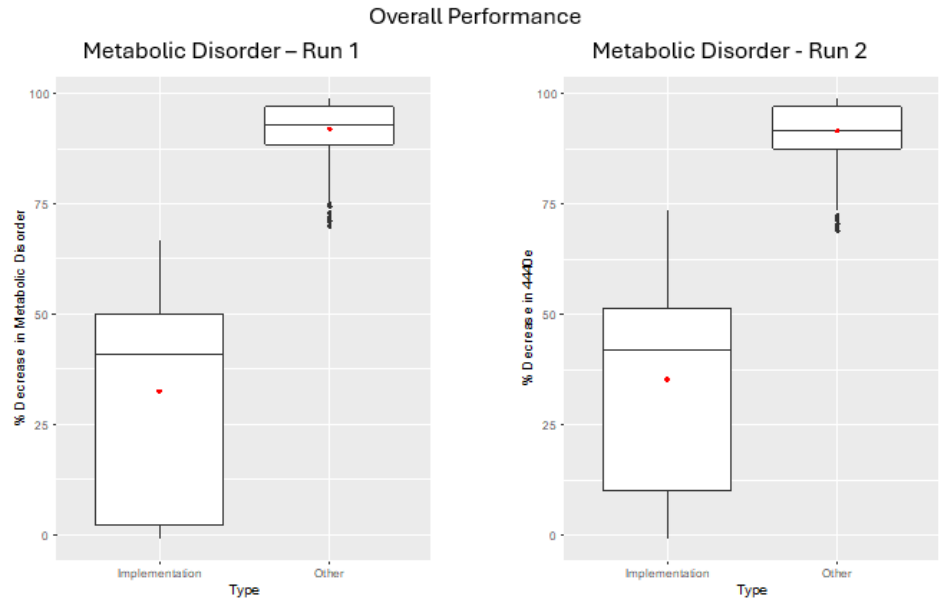
Importance Table – 00673-01-C-IPFQR Screening for Metabolic Disorders (SMD)

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Rate	0.15771	0.53855	0.72414	0.81576	0.87435	0.91469	0.94625	0.97131	0.99000	1.00000
Entities	144	145	145	144	145	145	144	145	145	145
Persons	52481	49727	50924	47182	47833	46989	47589	42778	37279	36889
Events	8,277	26,781	36,876	38,489	41,823	42,980	45,031	41,551	36,906	36,889
At benchmark	50,975	48,300	49,463	45,828	46,461	45,641	46,224	41,551	36,906	36,889
Person per entity	364	343	351	328	330	324	330	295	257	254

Rule of three Heuristic			NNTT	Decile Change
Importance (low bar)	64.02	MET	1.187	10/1 Persons
Importance (high bar)	5.72	MET	3.625	10/3 Persons
Overall		MET	317.7	Ave. Persons
			245.8	Ave. Events
			267.6	10/1 Events
			87.6	10/3 Events

- Importance table for PY2021
- Meets the “Rule of Three” heuristic
- NNTT is 3.6 persons
- Benchmark performance is 23.2%
- Counter-factual performance is 2.3%

Simulation – 00673-01-C-IPFQR Screening for Metabolic Disorders (SMD)



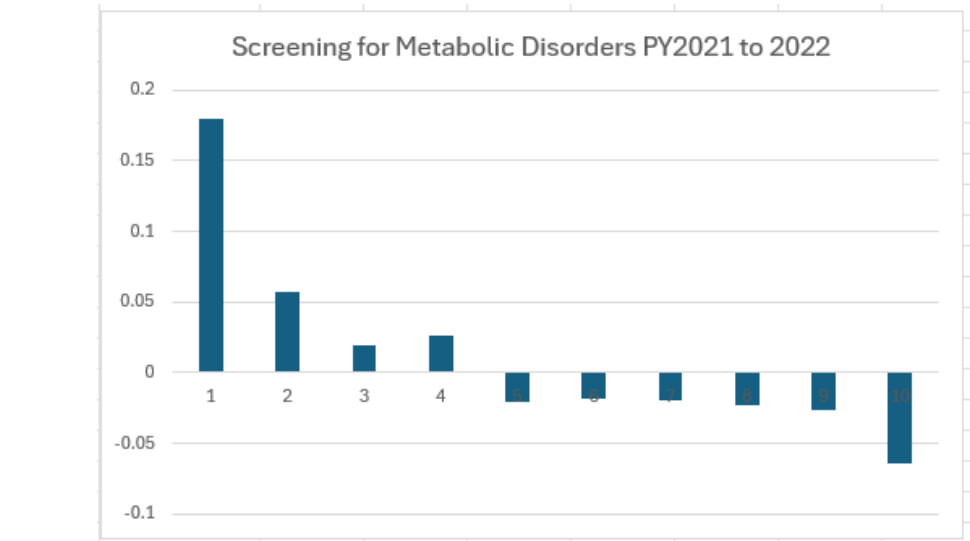
Variability is greater for the Implementation scenario compared to the Other scenario. Results are relatively consistent across the two sets of runs. For this metric, the improvement after achieving 80% overall implementation of the intervention is less than the impact of achieving the third decile.

		Mean	5%	25%	50%	75%	95%
Run 1	Implementation	32.48%	0.02%	2.31%	40.78%	50.02%	59.54%
	Other	91.82%	81.72%	88.24%	92.69%	97.06%	98.01%
Run2	Implementation	35.24%	0.16%	9.91%	42.08%	51.27%	59.74%
	Other	91.38%	83.08%	87.52%	91.34%	97.02%	97.93%

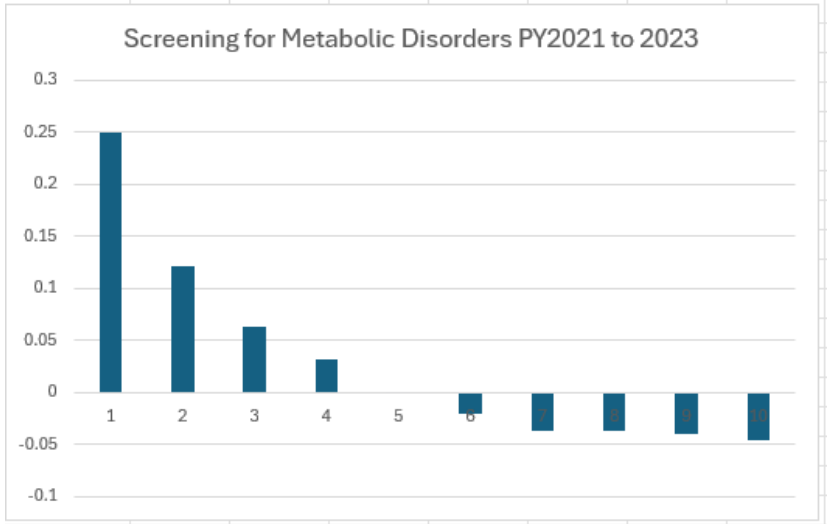
Simulation	LIB Benchmark	LB	UB	LIB Counterfactual	LB	UB
percent	0.92015	0.87880	0.97040	0.41430	0.06110	0.50645
exp()	0.39846	0.41528	0.37893	0.66080	0.94073	0.60263
base rate	0.22640	0.22640	0.22640	0.22640	0.22640	0.22640
new rate	0.09021	0.09402	0.08579	0.14960	0.21298	0.13643
LIB	-92.0%	-87.9%	-97.0%	-41.4%	-6.1%	-50.6%
base rate	0.77360	0.77360	0.77360	0.77360	0.77360	0.77360
new rate	0.90979	0.90598	0.91421	0.85040	0.78702	0.86357
HIB	16.2%	15.8%	16.7%	9.5%	1.7%	11.0%

- Benchmark performance (other) is 16.2% [15.8%-16.7%] increase (PY2021)
- Counterfactual (implementation) is 9.5% [1.7%-11.0%] increase (PY2021)

Performance Data – 00673-01-C-IPFQR Screening for Metabolic Disorders (SMD)



	2021			2022		
Entities	1,447			1,371		
Persons	459,671			460,984		
Events	355,603	0.7736		364,273	0.7902	
At benchmark	448,238	0.9751		449,691	0.9755	
Change	92,636	0.2015	23.2%	85,418	0.1853	21.1%
Events	355,603	0.7736		364,273	0.7902	
Plausible	363,886	0.7916		371,892	0.8067	
Change	8,283	0.01802	2.3%	7,620	0.01653	2.1%
%Performance Change						2.1%



	2021			2023		
Entities	1,447			1,388		
Persons	459,671			453,806		
Events	355,603	0.7736		364,680	0.8036	
At benchmark	448,238	0.9751		443,551	0.9774	
Change	92,636	0.2015	23.2%	78,871	0.1738	19.6%
Events	355,603	0.7736		364,680	0.8036	
Plausible	363,886	0.7916		371,217	0.8180	
Change	8,283	0.01802	2.3%	6,537	0.01440	1.8%
%Performance Change						3.8%

- From PY2021-2023, the counter-factual performance was 2.3%, 2.1%, and 4.4% (cumulative).
- The actual performance was 2.1%, 1.7%, and 3.8% (cumulative).

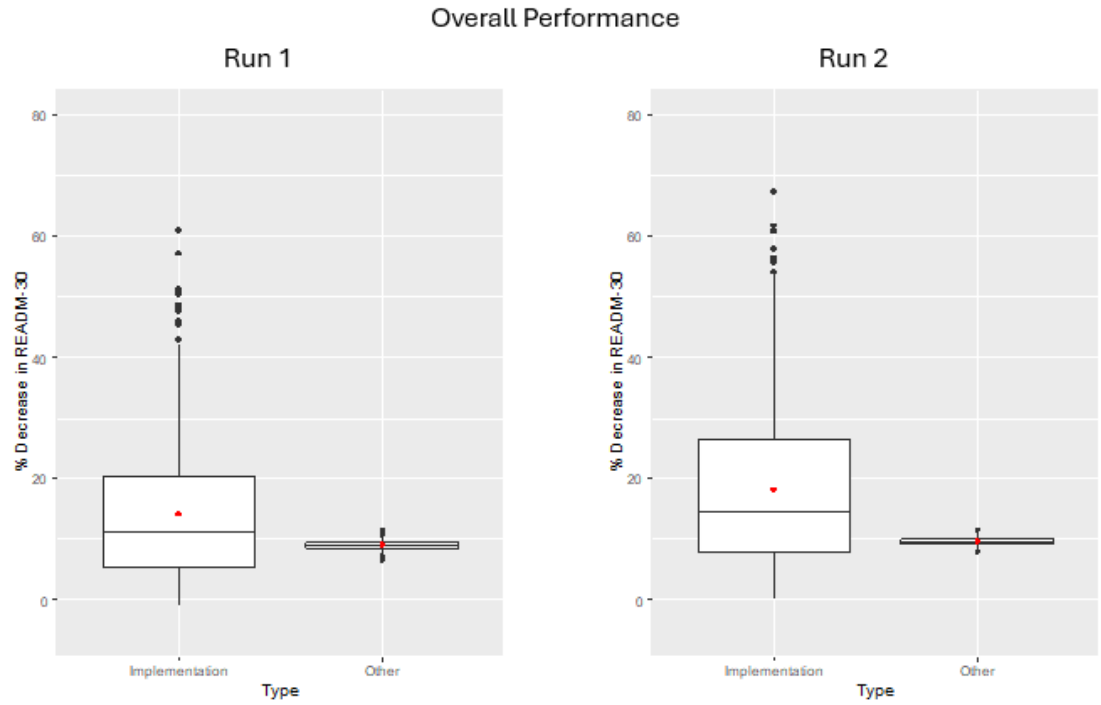
Importance Table – READM-30-IPF
Patients Readmitted to any Hospital within 30 Days of Discharge from the Inpatient Psychiatric Facility

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Rate	0.15487	0.16986	0.17916	0.18576	0.19152	0.19856	0.20560	0.21406	0.22477	0.25116
Entities	127	128	128	127	128	128	127	128	128	128
Persons	28012	26410	27824	32765	24419	29491	34083	30101	28720	38807
Events	4,338	4,486	4,985	6,087	4,677	5,856	7,007	6,444	6,455	9,747
At benchmark	4,338	4,486	4,985	5,870	4,375	5,284	6,106	5,393	5,146	6,953
Person per entity	221	206	217	258	191	230	268	235	224	303

Rule of three Heuristic			NNTT	Decile Change
Importance (low bar)	5.60	MET	10.384	10/1 Persons
Importance (high bar)	0.69	NOT	13.889	10/3 Persons
Overall		NMB	235.4	Ave. Persons
			47.0	Ave. Events
			22.7	10/1 Events
			17.0	10/3 Events

- Importance table for PY2022
- Does not meet the “Rule of Three” heuristic
- NNTT is 13.9 persons
- Benchmark performance is -12.7%
- Counter-factual performance is -1.5%

Simulation – READM-30-IPF
Patients Readmitted to any Hospital within 30 Days of Discharge from the Inpatient Psychiatric Facility

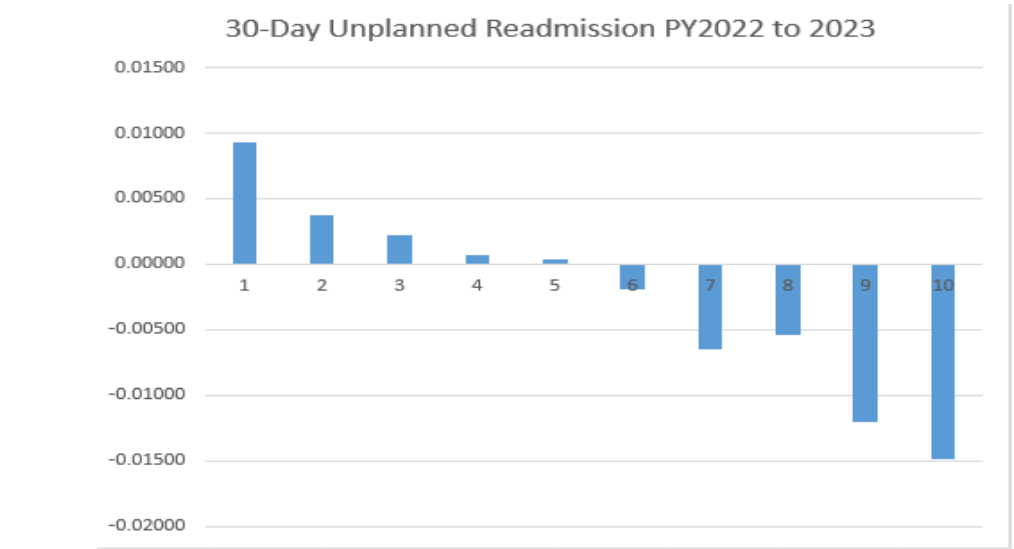


There is much more variability in the effectiveness of the 80% implementation scenario compared to the "Other" scenario. Both scenarios are consistent across the two sets of runs, with slightly higher impact of the 80% implementation scenario in the second set of runs. The 80% implementation scenario has a greater impact than the "Other" scenario.

		Mean	5%	25%	50%	75%	95%
Run 1	Implementation	14.28%	1.43%	5.41%	11.44%	20.40%	36.42%
	Other	9.01%	8.02%	8.60%	9.02%	9.39%	10.09%
Run2	Implementation	18.12%	2.60%	8.00%	14.64%	26.43%	42.74%
	Other	9.65%	8.67%	9.24%	9.68%	10.06%	10.63%

- Benchmark performance (other) is 9.3% [8.9% to 9.7%] decrease.
- Counterfactual performance (implementation) is 13.0% [6.7% to 23.4%] decrease.

Performance Data – READM-30-IPF
Patients Readmitted to any Hospital within 30 days of Discharge from the Inpatient Psychiatric Facility



	2022			2023		
Entities	1,277			1,264		
Persons	300,632			245,948		
Events	60,081	0.1998		48,464	0.1971	
At benchm	52,936	0.1761		42,640	0.1734	
Change	(7,145)	(0.0238)	-12.7%	(5,824)	(0.0237)	-12.8%
	Effect Size	0.12661		Effect Size	0.12803	
Events	60,081	0.1998		48,464	0.1971	
Plausible	59,202	0.1969		47,739	0.1941	
Change	(879)	(0.00292)	-1.5%	(725)	(0.00295)	-1.5%
%Performance Change						-1.4%

- From PY2022-2023, the counter-factual performance was -1.5%.
- From PY2022-2023, the actual performance was -1.4%.
- Facilities in deciles 1-5 (better) in PY2022 got worse.
- Facilities in deciles 6-10 (worse) in PY2022 got better.
- Regression to the mean: benchmark around deciles 5-6?

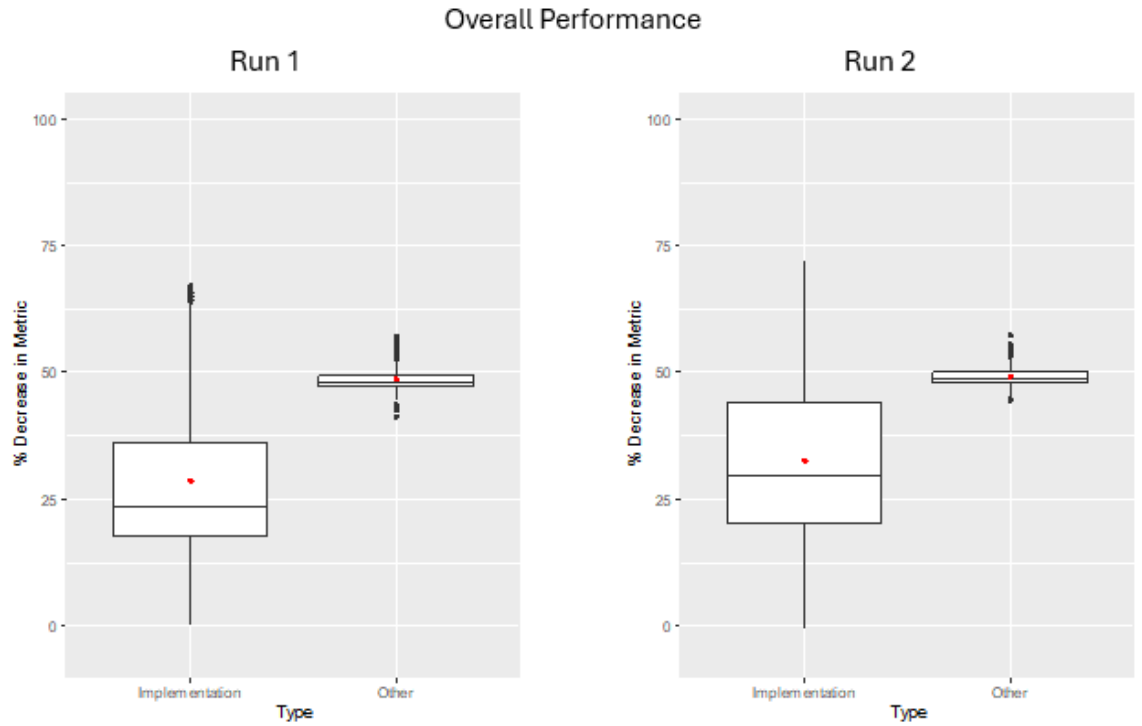
Importance Table – 00389-01-C-HHQR Influenza Immunization Received for Current Flu Season

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Rate	0.23622	0.48368	0.57453	0.63744	0.68400	0.72742	0.77124	0.82061	0.87904	0.96358
Entities	813	813	814	813	813	814	813	814	813	814
Persons	139710	284606	501624	567438	639596	736413	563588	445125	341315	226438
Events	33,002	137,658	288,198	361,708	437,483	535,679	434,663	365,274	300,029	218,190
At benchmark	114,647	233,550	411,637	465,645	524,859	604,308	462,486	365,274	300,029	218,190
Person per entity	172	350	616	698	787	905	693	547	420	278

Rule of three Heuristic			NNTT	Decile Change
Importance (low bar)	72.38	MET	1.37	1/10 Persons
Importance (high bar)	7.31	MET	1.71	1/8 Persons
Overall		MET	546.6	Ave. Persons
			382.6	Ave. Events
			397.6	1/10 Events
			319.4	1/8 Events

- Importance table for PY2022
- Meets the “Rule of Three” heuristic
- NNTT is 1.7 persons
- Benchmark performance is 17.3%
- Counterfactual performance is 1.9%

Simulation – 00389-01-C-HHQR Influenza Immunization Received for Current Flu Season



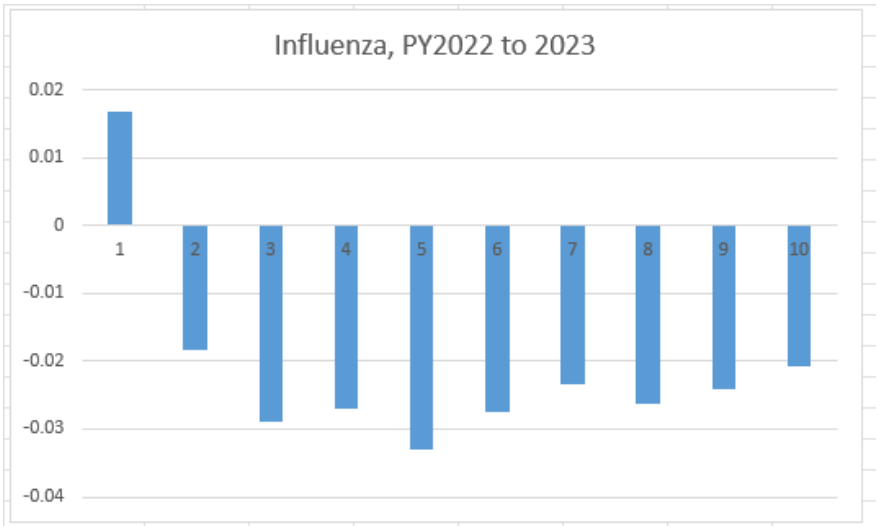
There is much more variability in the effectiveness of the 80% implementation scenario compared to the "Other" scenario. The 80% implementation scenario has a lesser impact than the "Other" scenario.

		Mean	5%	25%	50%	75%	95%
Run 1	Implementation	28.44	13.70	17.73	23.45	36.17	60.36
	Other	48.61	45.88	47.30	48.29	49.46	52.85
Run2	Implementation	32.52	12.37	20.27	29.63	44.17	61.92
	Other	49.08	46.27	47090	48.85	49.96	53.06

LIB Benchmark	LB	UB	LIB Counterfactual	LB	UB
0.48570	0.47195	0.49710	0.26540	0.19000	0.40170
0.61527	0.62378	0.60829	0.76690	0.82696	0.66918
0.30005	0.30005	0.30005	0.30005	0.30005	0.30005
0.18461	0.18717	0.18252	0.23011	0.24813	0.20079
-48.6%	-47.2%	-49.7%	-26.5%	-19.0%	-40.2%
0.69995	0.69995	0.69995	0.69995	0.69995	0.69995
0.81539	0.81283	0.81748	0.76989	0.75187	0.79921
15.3%	15.0%	15.5%	9.5%	7.2%	13.3%

- Benchmark performance is 15.3% [15.0%-15.5%] increase (PY2022).
- Counter-factual performance is 9.5% [7.2%-13.3%] increase (PY2022).

Performance Data – 00389-01-C-HHQR Influenza Immunization Received for Current Flu Season



Entities	8,134			8,104		
Persons	4,445,853			4,550,815		
Events	3,111,884	0.7000		3,061,745	0.6728	
At benchmark	3,700,625	0.8324		3,750,086	0.8240	
Change	588,741	0.1324	17.3%	688,341	0.1513	20.3%
	Effect Size	0.17327		Effect Size	0.20279	
Events	3,111,884	0.7000		3,061,745	0.6728	
At benchmark	3,171,361	0.7133		3,139,314	0.6898	
Change	59,477	0.01338	1.9%	77,570	0.01705	2.5%
%Performance Change						-4.0%

- From PY2022-2023, counterfactual performance was 1.9%,
- From PY2022-2023, the actual performance was -4.0%.
- Agencies in decile 1 (worse) in PY2022 got better.
- Agencies in deciles 2-10 (better) in PY2022 got worse.
- The performance change over 2 years was -5.1% (PY2021-2023).

Importance Table – 00199-01-C-MIPS Diabetes Mellitus: Diabetic Foot and Ankle Care, Peripheral Neuropathy–Neurological Evaluation (Clinician)

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Rate	0.03413	0.24844	0.61203	0.85141	0.97313	0.99813	1.00000	1.00000	1.00000	1.00000
Entities	63	64	64	64	64	64	64	64	64	65
Persons	16318	15155	19685	15884	15315	18891	18560	16896	21141	18091
Events	557	3,765	12,048	13,524	14,903	18,856	18,560	16,896	21,141	18,091
At benchmark	16,318	15,155	19,685	15,884	15,315	18,891	18,560	16,896	21,141	18,091
Person per entity	259	237	308	248	239	295	290	264	330	278

Rule of three Heuristic			NNTT	Decile Change
Importance (low bar)	58.74	MET	1.04	1/10 Persons
Importance (high bar)	3.50	MET	1.04	1/8 Persons
Overall		MET	274.9	Ave. Persons
			216.2	Ave. Events
			265.5	1/10 Events
			265.5	1/8 Events

- Importance table for PY2022 (scores not rates) – eligible clinician
- Meets the “Rule of Three” heuristic.
- NNTT is 1.0 persons.
- Benchmark performance is 24.0%.
- Counter-factual performance is 1.6%.

Importance Table – 00199-01-C-MIPS
Diabetes Mellitus: Diabetic Foot and Ankle Care, Peripheral Neuropathy – Neurological Evaluation (Group)

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Rate	0.12667	0.60871	0.89129	0.99000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Entities	30	31	31	30	31	31	30	31	31	31
Persons	76708	29831	33226	20575	15458	25866	19058	25075	24982	26859
Events	9,716	18,158	29,614	20,369	15,458	25,866	19,058	25,075	24,982	26,859
At benchmark	76,708	29,831	33,226	20,575	15,458	25,866	19,058	25,075	24,982	26,859
Person per entity	2,557	962	1,072	686	499	834	635	809	806	866

Rule of three Heuristic			NNTT	Decile Change
Importance (low bar)	268.67	MET	1.15	1/10 Persons
Importance (high bar)	16.80	MET	1.15	1/8 Persons
Overall		MET	969.5	Ave. Persons
			700.8	Ave. Events
			846.7	1/10 Events
			846.7	1/8 Events

- Importance table for PY2022 (scores not rates) – group
- Meets the “Rule of Three” heuristic.
- NNTT is 1.1 persons.
- Benchmark performance is 32.5%.
- Counter-factual performance is 2.4%.

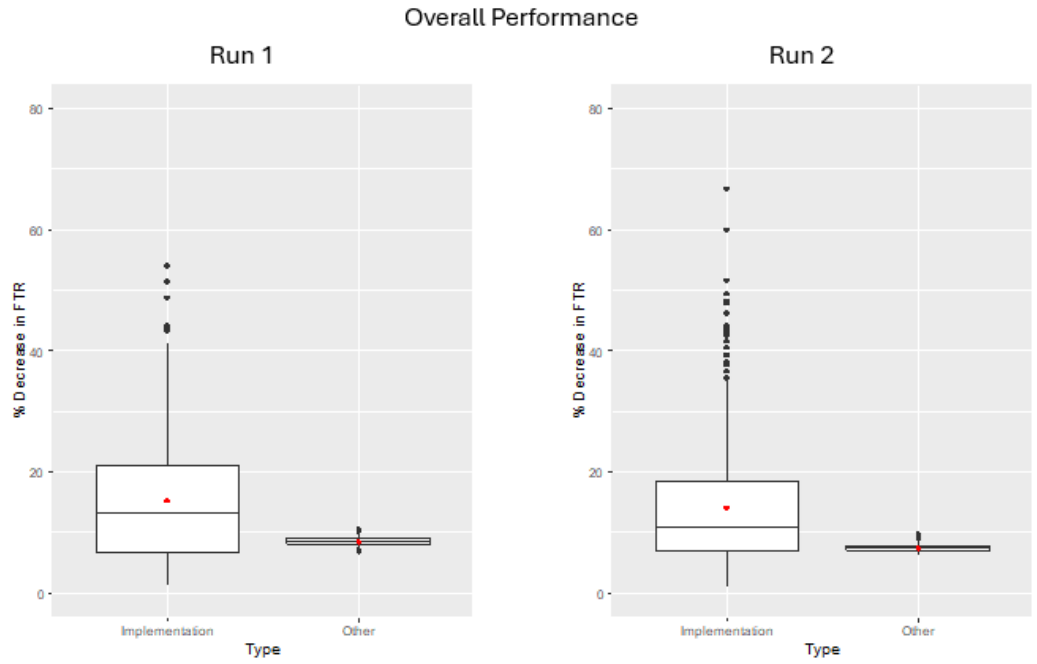
Importance Table – PSI-04
Death Rate Among Surgical Inpatients with Serious Treatable Complications

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Rate	0.1287	0.1421	0.1482	0.1524	0.1565	0.1607	0.1651	0.1705	0.1772	0.1928
Entities	153	154	154	154	154	154	154	154	154	155
Persons	18271	12564	12831	11703	12277	11953	15561	12571	14824	16887
Events	2,351	1,785	1,901	1,784	1,921	1,921	2,569	2,143	2,627	3,255
At benchmark	2,351	1,785	1,901	1,734	1,819	1,771	2,306	1,863	2,197	2,502
Person per entity	119	82	83	76	80	78	101	82	96	109

Rule of three Heuristic			NNTT	Decile Change
Importance (low bar)	1.32	NOT	15.597	1/10 Persons
Importance (high bar)	0.45	NOT	44.822	3/8 Persons
Overall		NOT	90.5	Ave. Persons
			14.5	Ave. Events
Effect Size	0.096		5.8	1/10 Events
			2.0	3/8 Events

- Importance table for PY2021
- Does not meet the “Rule of Three” heuristic.
- NNTT is 44.8 persons.
- Spreadsheet estimate:
  - Benchmark performance is -9.6%.
  - Counterfactual performance is –0.82%.

Simulation – PSI-04
Death Rate Among Surgical Inpatients with Serious Treatable Complications

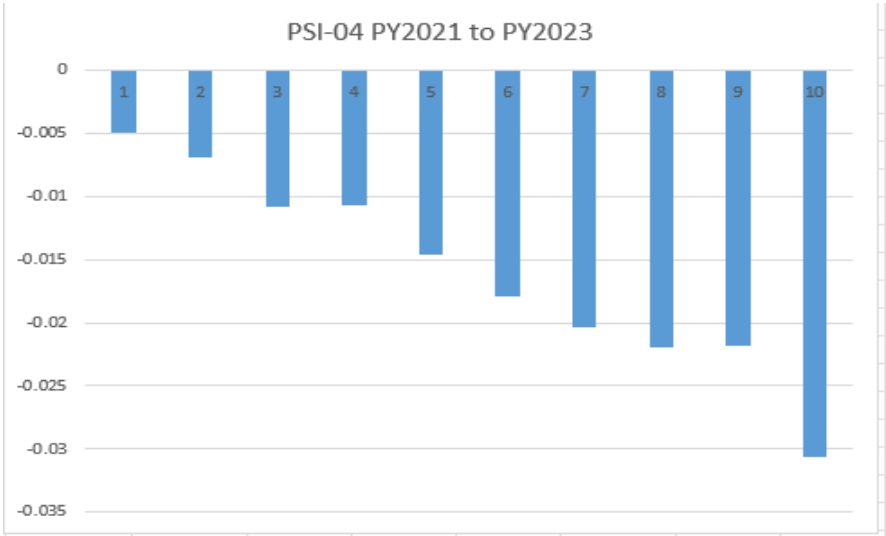


There is much more variability in the effectiveness of the 80% implementation scenario compared to the "Other" scenario. Both scenarios are consistent across the two sets of runs, with slightly higher impact of the 80% implementation scenario in the first set of runs. The 80% implementation scenario has a greater impact than the "Other" scenario.

		Mean	5%	25%	50%	75%	95%
Run 1	Implementation	15.17%	2.72%	6.74%	13.09%	21.05%	36.18%
	Other	8.54%	7.57%	8.15%	8.54%	8.92%	9.57%
Run2	Implementation	14.04%	2.98%	6.95%	10.90%	18.36%	34.48%
	Other	7.43%	6.47%	7.02%	7.41%	7.8%	8.43%

- Simulation estimate:
  - Benchmark estimate is 8.0% [7.6%-8.4%] (PY2021) decrease.
  - Counterfactual estimate is 12.0% [6.8%-19.7%] (PY2021) decrease.

Performance Data – PSI-04
Death Rate Among Surgical Inpatients with Serious Treatable Complications



	2021				2023		
Entities	1,540				1,432		
Persons	139,442				125,921		
Events	22,259	0.1596			17,995	0.1429	
At benchmark	20,230	0.1451			16,259	0.1291	
Change	(2,029)	(0.0145)	-9.6%		(1,736)	(0.0138)	-10.1%
	Effect Size	0.218			Effect Size	0.218	
Events	22,259	0.1596			17,995	0.1429	
At benchmark	20,613	0.1478			16,712	0.1327	
Change	(1,646)	(0.0118)	-7.7%		(1,284)	(0.0102)	-7.4%
%Performance Change							-11.06%

- From PY2021 to PY2023, the actual performance change was -11.1%.
- All facilities got better (benchmark?)
  - Specification change in PY2024
- Why was the counterfactual too low?
  - The *estimated* effect size (9.6%) is too low; the *meta-analysis* effect size (21.1%) is better but still low (-2.0%).
  - **Implies adoption-implementation 30-60%.**

Closing Thoughts

Evaluating Importance (Impact)

- The simulated counterfactual estimate is generally less than the benchmark estimate, but not always.
- The simple counterfactual estimate and simulation accounts for adoption, implementation, and intervention effectiveness.
- The counterfactual estimate in general aligns with the actual performance.
- The counterfactual estimate also provides a rebuttable presumption that may be informed by experience or evidence (literature).
- An informed effectiveness estimate infers adoption-implementation.
- The “Rule of Three” heuristic provides a rebuttable presumption of burden-benefit.

Use Impact Estimates for Measure Development and Implementation

Developers can use these metrics to:

- Understand the real-world potential of their measure and set realistic expectations for measure performance
  - The counterfactual estimate provides a plausible, evidence-informed projection.
  - Simulation allows developers to explore a range of scenarios based on varying adoption, implementation, and effectiveness rates.
  - The “rebuttable presumption” aspect allows developers to adjust expectations based on literature, context, or field experience.
- Prioritize measures with meaningful and achievable impact
  - The “Rule of Three” heuristic provides a simplified but evidence-informed view of burden-benefit trade-offs, helping focus on measures that offer the best return on investment.
- Evaluate whether a measure is worth scaling
  - This helps communicate the value proposition of the measure using credible, quantitative estimates.

Use Impact Estimates to Strengthen Endorsement and Maintenance Submissions

Strengthen the measure rationale by quantifying how the measure improves outcomes or reduces harm if entities perform at or above the benchmark.

Support logic models with realistic, evidence-informed projections of short- and long-term impact based on adoption, implementation, and effectiveness.

Demonstrate importance by showing measurable performance gaps and alignment between estimated and actual performance.

Communicate efficiency using metrics like the Number Needed to Treat (NNTT) and addressable event rates to highlight value and feasibility.

Anticipate impact by modeling potential gains, avoided adverse events, and burden-benefit tradeoffs to justify scale and spread.

Questions & Answers

References

- Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. Methods in Ecology and Evolution, 10(7), 949-959.
- Messick, S. (1989b). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan – articulation of “modern” validity theory
- Richman, B. D., Kaplan, R. S., Kohli, J., Purcell, D., Shah, M., Bonfrer, I., Golden, B., Hannam, R., Mitchell, W., Cehic, D., Crispin, G., & Schulman, K. A. (2022). Billing And Insurance-Related Administrative Costs: A Cross-National Analysis. Health affairs (Project Hope), 41(8), 1098–1106. https://doi.org/10.1377/hlthaff.2022.00241
- Saraswathula, A., Merck, S. J., Bai, G., Weston, C. M., Skinner, E. A., Taylor, A., Kachalia, A., Demski, R., Wu, A. W., & Berry, S. A. (2023). The Volume and Cost of Quality Metric Reporting. JAMA, 329(21), 1840–1847. https://doi.org/10.1001/jama.2023.7271
- https://en.wikipedia.org/wiki/Rule_of_three_(statistics)

References *(cntd., 1)*

- Shrank, W. H., Rogstad, T. L., & Parekh, N. (2019). Waste in the US Health Care System: Estimated Costs and Potential for Savings. JAMA, 322(15), 1501–1509. https://doi.org/10.1001/jama.2019.13978
- Standards for Educational and Psychological Testing. (2014). United States: American Educational Research Association – adoption of “modern” validity theory as a standard.
- Thissen D. (2001). Psychometric Engineering as Art. Psychometrika, 66(4), 473-485. doi:10.1007/BF02296190
- Tseng, P., Kaplan, R. S., Richman, B. D., Shah, M. A., & Schulman, K. A. (2018). Administrative Costs Associated With Physician Billing and Insurance-Related Activities at an Academic Health Care System. JAMA, 319(7), 691–697. https://doi.org/10.1001/jama.2017.19148
- Weissman, N. W., Allison, J. J., Kiefe, C. I., Farmer, R. M., Weaver, M. T., Williams, O. D., Child, I. G., Pemberton, J. H., Brown, K. C., & Baker, C. S. (1999). Achievable benchmarks of care: The ABCs of benchmarking. *Journal of evaluation in clinical practice*, *5*(3), 269–281. https://doi.org/10.1046/j.1365-2753.1999.00203.x

Partnership for Quality Measurement
Powered by Battelle

Presenting and Evaluating Cost Measures

Matthew Pickering | Battelle

Meridith Eastman | Battelle

October 15, 2025
2:10PM-3:10PM (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

Meridith Eastman | PRMR/MSR Task Lead

- Leads strategic and technical activities for PRMR/MSR
- 18 years public health and quality experience

E&M: Endorsement and Maintenance PRMR: Pre-Rulemaking Measure Review MSR: Measure Set Review

Session Objectives and Agenda

- Objectives
  - Review key challenges in the development and evaluation of cost measures
  - Discuss strategies for interpreting cost within the context of quality care
- Agenda
  - Overview of Cost Measures
  - Cost Measure Challenges and Solutions
  - Cost Measures and the E&M Evaluation Criteria
  - Resources
  - Q&A

Overview of Cost Measures

Overview

What is a cost measure?

- Cost measures estimate the costs of health care services from a payer perspective.
- Cost measures may be:
  - Episode based, covering a range of procedures, acute inpatient medical conditions, and chronic conditions.
  - Population based, focusing more broadly on the entire spectrum of care.
- Cost measures are not a stand-alone quality construct but rather must be interpreted in the context of quality to capture efficiency or value (outcome/cost).

Overview

Why do we have cost measures?

- **Clinician's role:** While clinicians do not determine the price of individual services provided to Medicare patients, they can affect the amount and types of services provided. By better coordinating care and seeking to improve health outcomes, clinicians play a meaningful role in delivering high-quality care at a reasonable cost.
- **Value:** Cost and quality are not mutually exclusive. In fact, when considered together, quality and cost represent health care *value*. Health care value is the measured improvement in a person's health outcomes for the cost of achieving that improvement.
- **Requirements:** The Merit-based Incentive Payment System (MIPS) must include measures of cost according to Section 1848(r) of the Social Security Act which was added by section 101(f) of the Medicare Access CHIP Reauthorization Act (MACRA).

Overview

Potential benefits and harms of cost measures

- While cost measures can promote efficiency, there is a perceived risk of limiting necessary care or discouraging treatment of high-cost patients.

Primary Benefits	Primary Harms
• More efficient care (outcome/cost) • Fewer complications and/or avoidable utilization (e.g., ED visits, additional procedures, inpatient admissions or readmissions)	• Potential to drive down utilization and decrease access to necessary care (i.e., stinting) • Potential to discourage treatment perceived to be high cost

Cost Measure Challenges and Solutions

Cost Measure Challenges

- Evaluating cost measures can be challenging due to:
  - Complexity of the specification.
  - Difficulty in attributing costs to clinicians or groups.
  - Potential relationship between better cost performance and quality of care
- Following the guidance in this presentation will help facilitate E&M and PRMR review of cost measures.
  - This guidance is not required, but it is recommended.

Cost Measure Challenges and Potential Solutions

Problem	Potential Solution
Standardized prices or dollar-weighted services do not reflect the resources available for service delivery (i.e., actual resource use and costs may differ from estimates based on standardized prices).	Stratify entities by payer mix (e.g., Medicare, Medicaid, commercial, other)
Potential for a decrease in necessary care	Identify services that may be substituted with lower quality care, and stratify entities based on the propensity to receive those services (post risk adjustment). Consider whether valuable higher cost services (e.g., rehabilitation services) are being used at appropriate levels for appropriate patients and whether risk adjustment or stratification adequately protects against this. OR Consider excluding services in which there is a high likelihood of stinting on a specific service due to high cost.
Demonstrate more efficient care (outcome/cost), as shown in the green boxes in the next slide (i.e., better quality performance for a given level of cost performance or, equivalently, better cost performance for a given level of quality performance)	Populate the relationship between better performance on the cost measure and better performance on relevant measures of quality (process or outcome)

Cost Measure Challenges and Potential Solutions *(cntd., 1)*

	Quality Performance		
Cost Performance	**Worse**	**Neutral**	**Better**
Better (i.e., Lower Cost)	Somewhat efficient	Efficient care	Efficient care
Neutral	Not efficient care	Somewhat efficient	Efficient care
Worse (i.e., Higher Cost)	Not efficient care	Not efficient care	Somewhat efficient

Cost Measure Challenges and Potential Solutions *(cntd., 2)*

Figure 1. Elective Hip Arthroplasty Episode-Based Cost Measure Logic Model

Inputs	Activities	Outputs	Outcomes	Impacts
• Multidisciplinary care team (surgeons, nurses, physical therapists, care coordinators) • Evidence-based clinical guidelines for hip arthroplasty • Data systems for tracking costs and outcomes • Patient education materials • Resources for care coordination • Financial and administrative support	• Implement standardized, evidence-based care pathways for elective hip arthroplasty • Provide preoperative patient education and optimization (e.g., managing comorbidities, smoking cessation) • Coordinate discharge planning and postoperative care (e.g., physical therapy, follow-up visits) • Monitor and analyze episode costs and outcomes • Engage in continuous quality improvement (CQI) based on data feedback	• Increased use of standardized care protocols • Enhanced patient engagement and preparedness • Improved coordination of care transitions • Regular reporting and review of cost and outcome data • Identification of areas for improvement	**Short Term:** • Reduced variation in care delivery • Improved patient understanding and engagement • Timely discharge and follow-up care • Early identification of inefficiencies and cost drivers **Intermediate Term:** • Lower rates of complications and readmissions • More efficient use of resources (e.g., appropriate post-acute care utilization) • Decreased episode costs for hip arthroplasty • Improved clinician performance on measure **Long Term:** • Sustained reduction in total cost of care for elective hip arthroplasty • Improved patient functional outcomes and satisfaction • Widespread adoption of best practices across entities • Enhanced value and efficiency in orthopedic care	• System-wide improvements in the value and affordability of elective orthopedic procedures • Greater sustainability of Medicare and improved access to high-value care for beneficiaries • Reduced financial burden for patients and payers

Cost Measure Challenges and Potential Solutions *(cntd., 3)*

Problem	Potential Solution
Demonstrate fewer complications or avoidable utilization (e.g., emergency department visits, additional procedures, and inpatient admissions or readmissions)	Association and mechanistic studies on better performance on the cost measure and lower rates of complications and avoidable utilization. Focus on care that most clearly reflects complications or avoidable care (e.g., emergency department use, higher-than-expected numbers of provider/facility visits).
Attribution model excludes or erroneously includes important specialists	Obtain early input from the technical expert panel (TEP) while the attribution model is developed. Establish face validity of the attribution model. Articulate clearly which clinicians are included (or excluded from the attribution model and why.

Cost Measures and the E&M Evaluation Criteria

Importance

Rubric Domain	Guidance

Importance	• Cost measures should be developed within the framework of a logic model to ensure a structured approach that clearly links them to health care priorities related to efficiency and value (i.e., improving outcomes while reducing costs). Priorities may include:
	  • Addressing the leading causes of inefficient care as a driver of morbidity and mortality
	  • Targeting areas of high resource use
	  • Focusing on conditions with high severity
	  • Addressing financial strain and burden experienced by patients due to high out-of-pocket health care costs
	• To ensure relevance and effectiveness in tackling critical health challenges, there must be a clear, evidence-based description of the resources and actionable steps that an accountable entity can adopt to improve cost measure performance. This ultimately enhances care quality and ensures patient safety.

Scientific Acceptability

Rubric Domain	Guidance
Reliability	• Reliability testing should ensure that the measure produces consistent results. This consistency is crucial for safety, as it guarantees that the measure reliably identifies the intended health care quality aspects without variation that could lead to misinterpretation or errors.
Validity	• Validity of cost measures refers to an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the extent to which a measure accurately captures the intended costs that are influenced by the accountable entities attributed to those costs. The measure focus is on standardized prices, ensuring that the measure reflects reimbursements rather than the cost of providing those services.

Scientific Acceptability *(cntd., 1)*

Rubric Domain	Guidance
Validity (cont.)	• Validity of cost measures involves ensuring that the measure can be impacted by those accountable entities through the reduction of the billable utilization of services and properly adjusts for external factors not under their influence. This means that there is evidence of accountable entities' ability to impact billable services (e.g., mechanistic studies), which is essential for the measure's effectiveness. This supporting evidence should include an explicit consideration of treatment choices that can influence costs. By incorporating standardized measures like Hierarchical Condition Categories (HCCs), clinical severity measures (as applicable), and appropriate treatment indicators into risk adjustment covariates, there is a better account for bias, ensuring more accurate assessments of accountable entities' performance. • Additionally, establishing acceptable model performance may be achieved by benchmarking risk adjustment models against previous measures with similar populations.

Feasibility and Usability

Rubric Domain	Guidance
Feasibility	• Feasibility of cost measures involves ensuring that the required data are readily available or can be captured without undue burden, thus facilitating their implementation for performance measurement. This ensures that the measure can be implemented effectively without imposing excessive burdens on health care providers, ultimately leading to more informed decision-making and improvements in care delivery.

Usability	• Usability of cost measures focuses on their practical application and impact on health care decision-making. This involves:
	  • Tracking progress and identifying areas for improvement
	  • Actively seeking and incorporating feedback from accountable entities and other interested parties to refine the measure and enhance its effectiveness
	• Usability also emphasizes the importance of ongoing improvement in results and the identification of unexpected findings, both positive and negative, to ensure the measure remains relevant and beneficial over time.

Usability *(cntd., 1)*

Rubric Domain	Guidance

Usability (cont.)	• While a measure may be valid, poorly designed incentives can lead to negative outcomes, such as reducing the quality of care to cut costs. This highlights that unintended consequences are not necessarily a threat to the validity of a measure but should be considered in the measure's use and usability.
	• In terms of the potential decrease in necessary care, this refers to the risk that cost measures might inadvertently lead to reductions in essential health care services as entities aim to lower costs.
	  • Addressing this involves pinpointing specific health care services that might be replaced by lower-quality alternatives, which could compromise patient care. Then, after risk adjustment, stratifying entities based on their propensity to provide those services.
	• To effectively analyze and interpret the relationship between cost performance and quality of care, Slide 11 provides an example framework for cross-referencing the Importance Table (performance score by decile) at the entity level.
	  • This involves comparing the cost measure with one or more relevant measures of process or outcome.

Cost Measure Submissions Best Practices

Ground the Measure in a Logic Model

- Use a structured logic model to link cost measurement to health care priorities such as efficiency, value, and patient outcomes.
- Clearly define inputs, processes (actions), outputs, and outcomes.

Provide clear interpretation of results and rationale for methodological choices

Address standardized pricing limitations

- Acknowledge that standardized prices may not reflect actual resource use.
- Stratify entities by payer mix or service substitution risk to improve fairness.

Articulate cost-quality impact

- Show how better cost performance correlates with better quality (process or outcome) performance.
- Use stratification or exclusions to protect against stinting on necessary care.

Address unintended consequences

- Identify services at risk of underuse due to cost pressures.
- Use exclusions or stratification to safeguard care quality.

Resources

- E&M Guidebook
  - Appendix H6 contains CBE Guidance on Cost Measures.
- What Good Looks Like – Cost Measure Example
  - Illustrative example responses to items in the full measure submission form.
- Technical Assistance
  - Email PQMsupport@battelle.org.

Questions & Answers

Thank You For Attending Day 1!

- Join us tomorrow for day 2 of the workshop!
- Thursday, October 16 | 12:00 PM-3:30 PM ET
- What to Expect:
  - **Applying Reliability Methods:** Review reliability methods for quality measures, focusing on appropriate use by data type and use case, with discussion of common challenges, tradeoffs, and mitigation strategies for low reliability.
  - **Validity Best Practices***:* Review best practices for entity-level testing, focusing on hypothesized relationships between measures, expected correlations, and using mechanistic evidence to support validation per CBE requirements.
  - **Risk Adjustment: Variability and Innovation:** Examine risk adjustment practices and challenges at the entity level, explore emerging tools such as machine learning, and provide feedback on opportunities for standardization (e.g., c-statistics, benchmarks).

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.