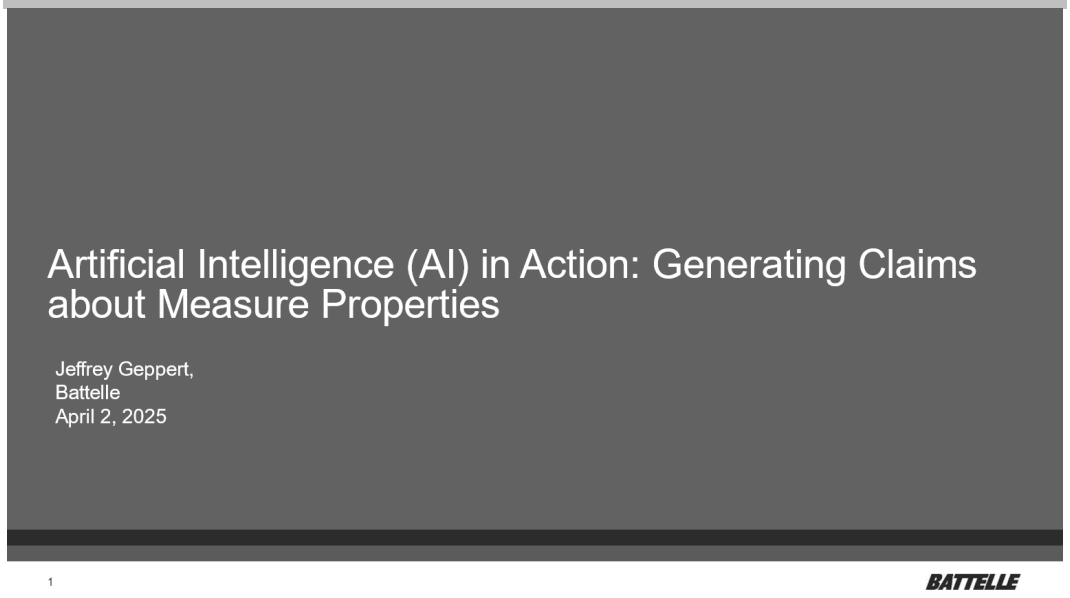
April Information Session Transcript

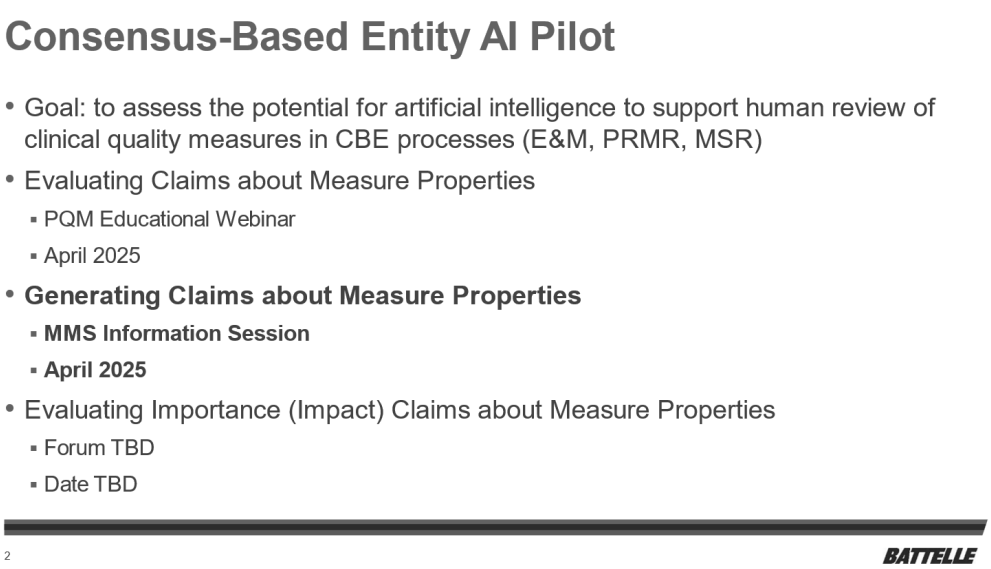
[SLIDE 1]



MODERATOR: Good afternoon, everyone. Thank you for joining this MMS Information Session. The topic for today is "Artificial Intelligence (AI) in Action: Generating Claims About Measure Properties." Just a reminder, at the end of the session there will be a Q&A. We'll try to get to as many as we can by the end. So today's presenter is Jeffrey Geppert, and we'll turn it right over to him to get started.

GEPPERT: Thank you. Good day, everyone. Thank you for joining. We're grateful for your participation. So today we're going to be discussing the use of artificial intelligence (AI) and the development and evaluation of clinical quality measures (CQMs). We refer to this work as the AI Pilot. So I'm Jeff Geppert. I'm the measurement science team lead for the consensus-based entity (CBE) at Battelle.

[SLIDE 2]

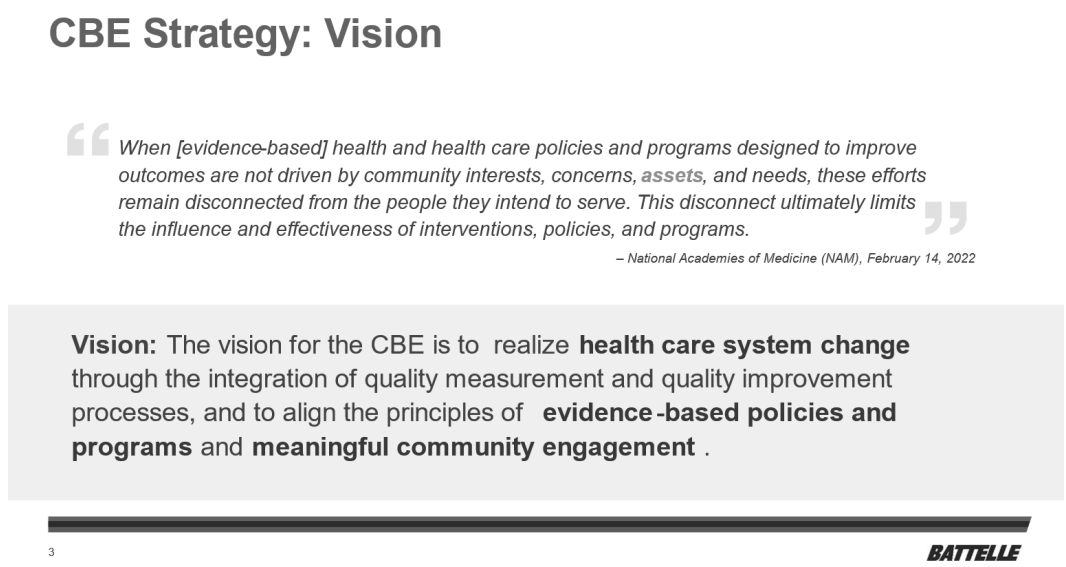


GEPPERT: So our main takeaway for today is that AI is really more than just performing measure development and evaluation tasks in a faster or automated or sort of at-scale manner. Rather, AI really does have the potential to transform the nature of those development and evaluation tasks, and it does so in three ways. Sort of used, you know, judiciously, AI has the potential to advance really any scientific domain, but clinical quality measurement specifically into a more predictable and mechanism-based and evidence-integrated scientific discipline, for the purposes of improving quality and safety, reducing burden, and reducing avoidable utilization.

So we've scheduled three webinars to discuss these three topics of predictable mechanism-based and evidence-integrated. So today's topic focuses on making measure development and evaluation more

"mechanism-based" using large language models (LLMs) to generate claims about measure properties. And then a PQM education webinar in a few weeks will focus on making development and evaluation more "evidence-integrated." We'll then follow up with a future webinar that will focus on making measure development and evaluation more predictable.

[SLIDE 3]



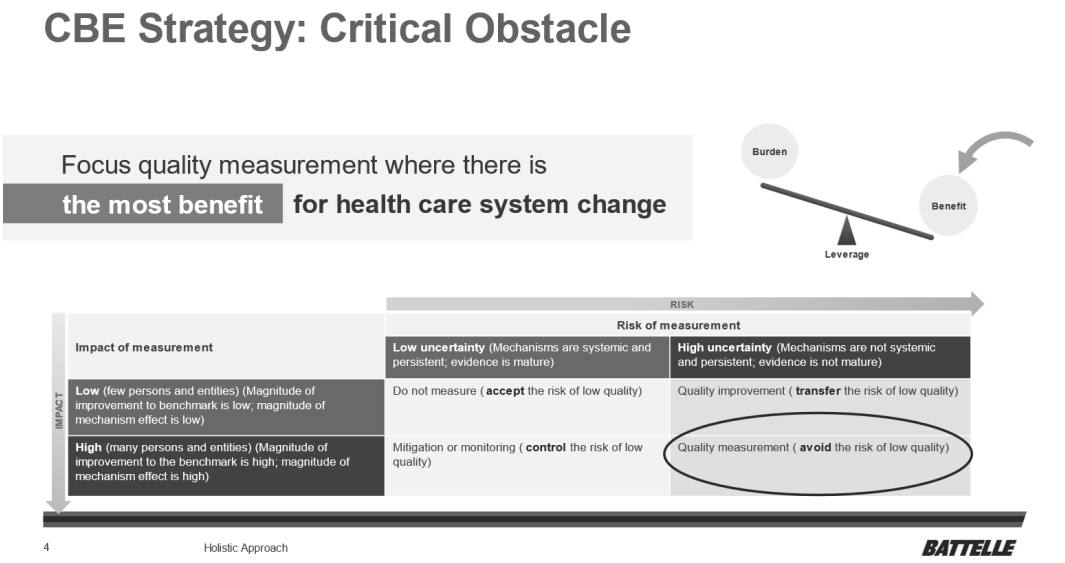
GEPPERT: So we always start with our vision. We won't talk too much about sort of our vision and strategy. We did have some webinars that we hosted in the fall available on the PQM website that talk more about sort of our strategy and approach, and we encourage you to look at those.

The two things to highlight here is that we're really focused on sort of a transformational change — healthcare to system change — in how quality measurement and more specifically these CBE processes to again support that transformational change, and we do so through both using

evidence and a meaningful community engagement. We really feel like those two things are mutually supportive, you know, to get more and better evidence, better interpretation of evidence, and engaging with the community.

Conversely, the community engagement is a more grounded and shared experience when it's evidence-based. So those two principles kind of simultaneously, and that really guides our thinking about AI as well.

[SLIDE 4]



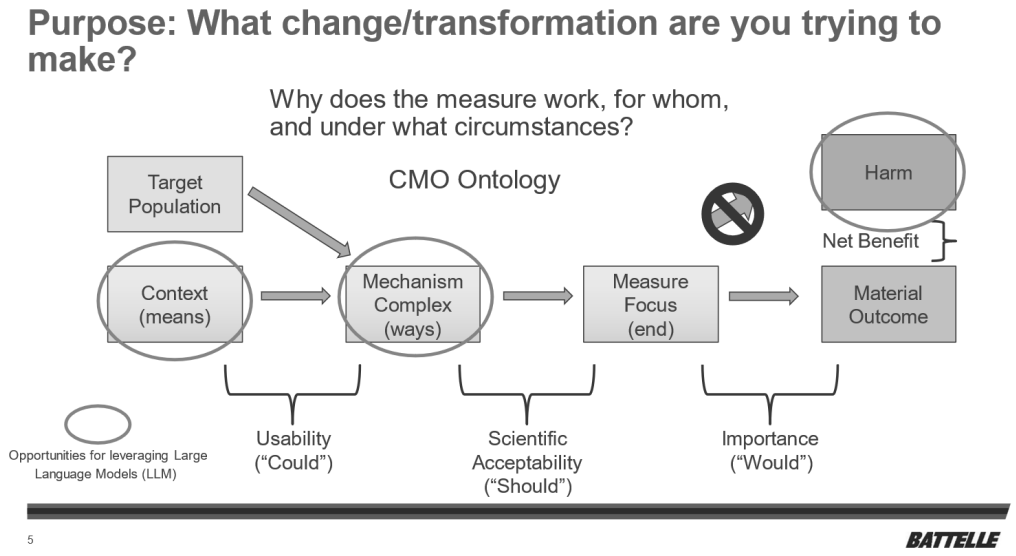
GEPPERT: So the specific problem we're trying to address is around the perceived burden of quality measurement and how we can use CBE processes to both reduce the burden and enhance the benefit of quality measurement. We think about that in terms of focusing quality measurement resources in this kind of lower right-hand box where both there's a high impact of quality measurement, so that the use of a quality

measure will result in a large decrease in adverse events or an increase in positive events.

We know how to do that, and the effect of that is high, but also that there's a great learning opportunity for quality measurement, sort of a "sweet spot." We know in general how to improve quality, but there's still some uncertainty that we can use quality measurement to learn about and so we want to do two things.

We want to be able to identify what those barriers are preventing clinicians and facilities from improving and to mitigate those barriers. And then we also want to identify what it is that better-performing clinicians and entities are doing, and then try to emulate those and create resources that allow everyone to emulate those. So that's kind of our strategy is to try to focus quality measurement kind of in that sweet spot.

[SLIDE 5]



GEPPERT: So this is kind of the structure that we use for evaluating quality measurement. Again, we're focused on what the purpose of the measure is, and what change or transformation is the measure trying to bring about. And then we use this, what's called a "realism" framework to really dig deeper and to understand exactly how or why does the measure actually work, for whom does it work, and under what circumstances does it work. So to try to uncover this we use this ontology which is called a "context mechanism outcome ontology."

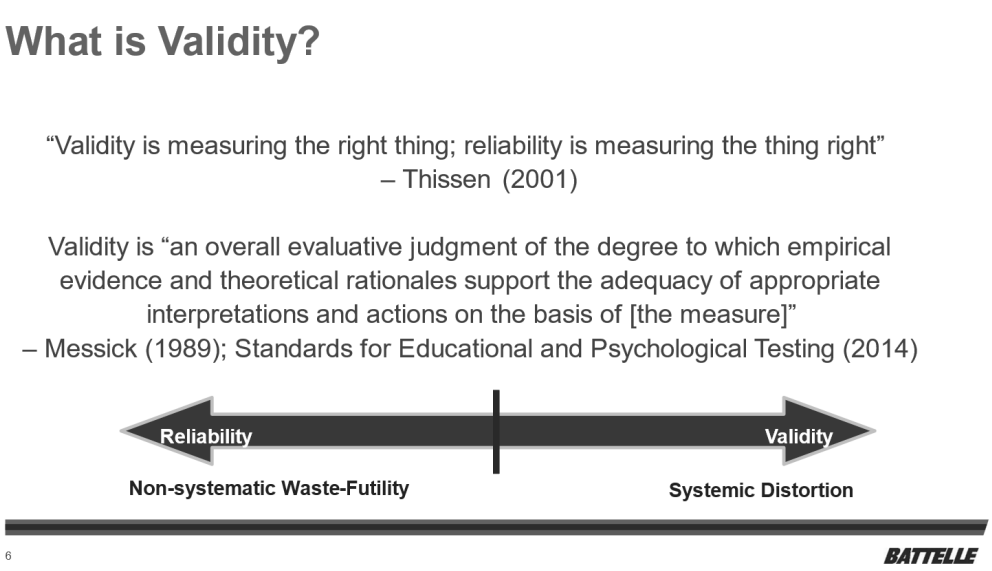
So historically, measurement evaluation is focused on that connection between the measure focus and material outcome, which we call the "importance consideration." Is this measure focus related to something that matters? The evidence that's been brought forth has largely been in support of that sort of relationship. What has been less common in measure evaluations is focus on the mechanisms necessary to bring that measure focus about

So we call claims that are associated with what it is that clinicians and facilities need to do in order to improve performance on the measure. We call that our "scientific acceptability" focus. And then more broadly the context in which those mechanisms are used and whether the barriers or the facilitators to using those mechanisms, and we refer to that as sort of our "usability evaluation." Finally, measure evaluation is typically focused on sort of the positive outcomes, and they're less specific about the potential harms associated with measure use.

So one of the goals of the AI Pilot was to see if we could use large language models (LLMs), which are very good at being contextually

appropriate in their responses, and if we could use them to address some of these historically underutilized areas of focus. We'll see some examples of that as we kind of walk through this today.

[SLIDE 6]



GEPPERT: So I want to take a moment to go a little bit more into what we think of as the validity or scientific acceptability of a measure. We use this definition which comes from Samuel Messick, and which has been adopted as the standard of validity in the educational and psychological testing area. So the definition of validity, as you can see here, 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."

So there are four elements of this that I want to emphasize. One is that validity is a *judgment*, and so it's a decision. We want to bring as much light as we can to that decision. So we want to be as explicit as we can about what are the claims, what is the evidence in support of those claims,

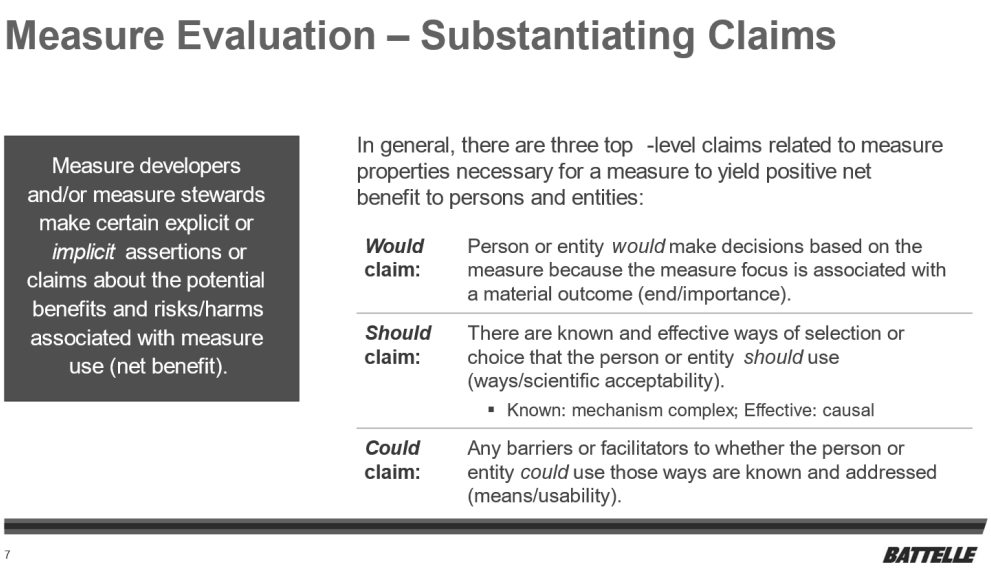
what are the arguments in support of those claims, and we want to be as explicit about those things as we can to inform that decision-making.

Validity is a matter of degree, and so there's always some kind of a residual risk associated with these decisions, and validity demonstrations are largely about ruling out sort of other potential causes of a measure focus — other than the quality of care. But then the mechanism focus, you know, changes that a little bit and makes the demonstration of validity sort of a more direct thing.

Evidence in theory, and so we look at evidence. We look at empirical studies, but we also want a very sort of theory about what clinicians and facilities choose to do to improve performance on a particular measure. And then that theory needs to be *plausible*. It needs to be logical, and we want to make sure that that theory is logical and plausible.

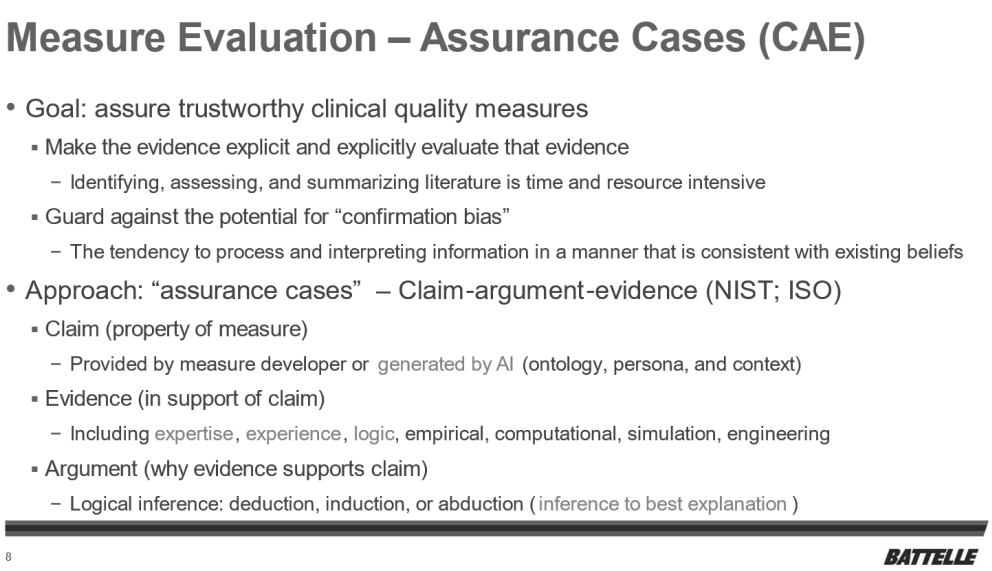
And then finally, interpretations and actions. So some people call this a response-based validity, argument-based validity. Again, it's looking at sort of what it is that clinicians and facilities do in response to a measure. What are the choices that they make? What are the resources they have available that determines the degree to which the measure is valid?

[SLIDE 7]



GEPPERT: So we talk about claims. Again, going back to the CMO ontology we're talking about "would, should, and could." A person "would" make decisions based on the measure, because it's associated with the material outcome. "Should," there are known and effective ways of selection and choice, that the person or entity should use. And then "could," any barriers or facilitators to whether the person or entity could use those ways are known and addressed. So those are the claims that we're going to be evaluating and generating using these large language models (LLMs).

[SLIDE 8]



GEPPERT: So our approach to evaluating those claims is to use what's known as an "assurance case." An assurance case consists of these three components — claim-argument-evidence (CAE). Now, assurance cases have been around for a while. They're a NIST and ISO standard common in the system safety sort of literature, but the benefit of AI is that it makes it much more possible, feasible, to actually build these claim-argument-evidence (CAE) assurance cases, and to maintain them over time to sort of build up an evidence base for the measure over time.

So our goal with these CBE processes, and with the AI Pilot in general, is to make sure that the quality measures are trustworthy, and we do that in two ways. One is that we want to make all the evidence explicit and explicitly evaluate that evidence that's about the judgment part; however,

identifying, assessing, summarizing literature is very time and resource-intensive. AI can potentially help with that.

Second, we want to guard against the potential for "confirmation bias," which is this tendency to process and interpret information in a matter that's consistent with beliefs. So this would be like "I believe the measure is valid, and so I'm going to be looking for evidence that supports the validity of the measure." We want to make sure that we're identifying and evaluating evidence that may, you know, be contrary to that claim.

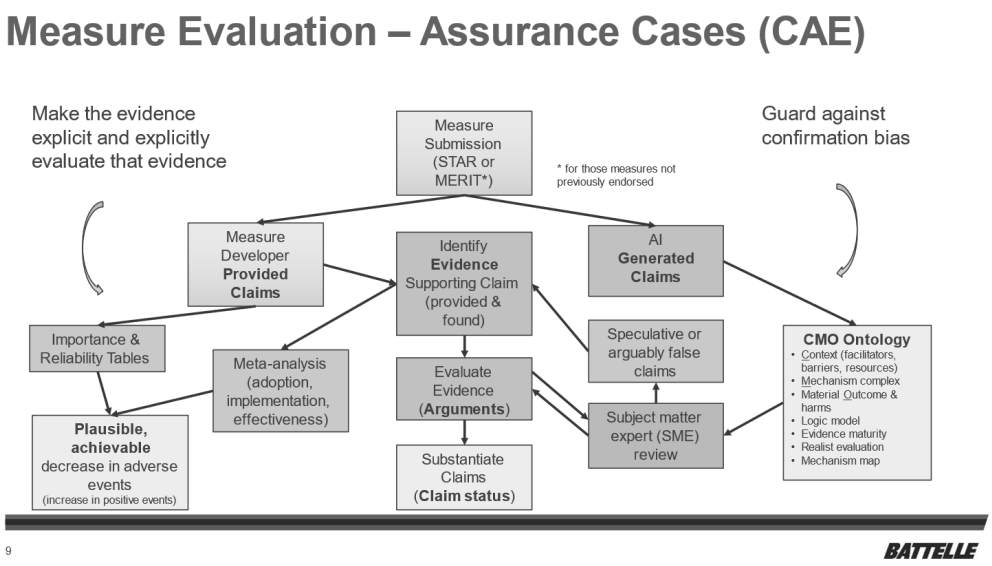
So assurance cases have these three components. They have the "claim" which is a property of the measure, and we have two sources of claims. We have claims that are provided by the developer, and then we're looking at claims that are generated by AI.

Second, we have evidence in support of the claim. Assurance cases take a very broad view of evidence, and so they incorporate both experience and expertise and logic, which tends to kind of be very important in terms of assessing the claims that are generated by AI.

In addition to empirical studies, computational approaches, simulation approaches, engineering approaches — all of these things which contribute to a sort of substantiating the truth of a claim. And then the part that is often missing is the argument part, which is why exactly the evidence supports the claim. These arguments are logical inferences. They're either deductive arguments, inductive arguments, or abduction arguments.

Abduction is sometimes referred to as “inference to best explanation (IBE).” So that’s a type of argument where you observe the effect, and then you sort of use logic to infer the cause. That’s the kind of argument we’re typically involved in with quality measures where we observe the results, and we’re trying to infer that the result is caused by the quality of the entity and not something else. So that’s the type of explanation that we’re typically trying to utilize. AI can help us sort of formulate those arguments as well as evaluate the evidence.

[SLIDE 9]



GEPPERT: So I won’t spend too much on this, but this is kind of how the whole system is put together, and the three webinars that I mentioned earlier are sort of going to be addressing all three components of this graphic. What we’re going to be focused on today is this right-hand side where we start with the measure submission that has a numerator and a denominator. We generate claims from AI about that measure focus and

target population. We use our CMO ontology to help structure those claims, and then we conduct a series of analysis. And then the intention is to have a subject matter expert (SME) review those claims, and then determine whether the claims are arguably true or potentially arguably false or speculative. And if they are arguably speculative or arguably false, then it feeds back into the evidence system where we look for published literature that supports the claims, or does not support the claims. And then we evaluate it in the way that will be described more in our webinar in a few weeks.

[SLIDE 10]

Measure Evaluation – Claim and Sub-claim Types

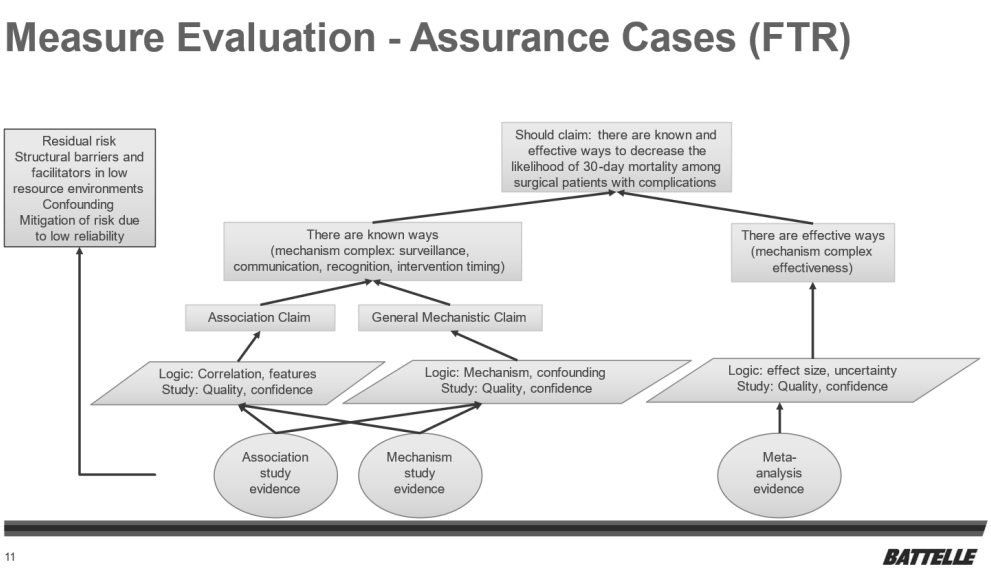
Importance	Would claim: Person or entity would make decisions based on the measure because the measure focus is associated with a material outcome
Validity	Should claim: There are known and effective ways of selection and choice that the person or entity should use
-Association	There is an association between the person or entity response to the measure and the measure focus
-Mechanism	There is an explicit articulation of the mechanisms (resources and response to those resources) responsible for the association
Usability	Could claim: Any barriers or facilitators to whether the person or entity could use those ways are known and addressed

10 BATTELLE

GEPPERT: So that's sort of the big picture of what we're talking about here. The point here is just to note that there are really two types of validity claims that we're looking at. There's what's called an "association claim," which is kind of what we typically think of when we're trying to show a measure is valid. We look at associations between two measures,

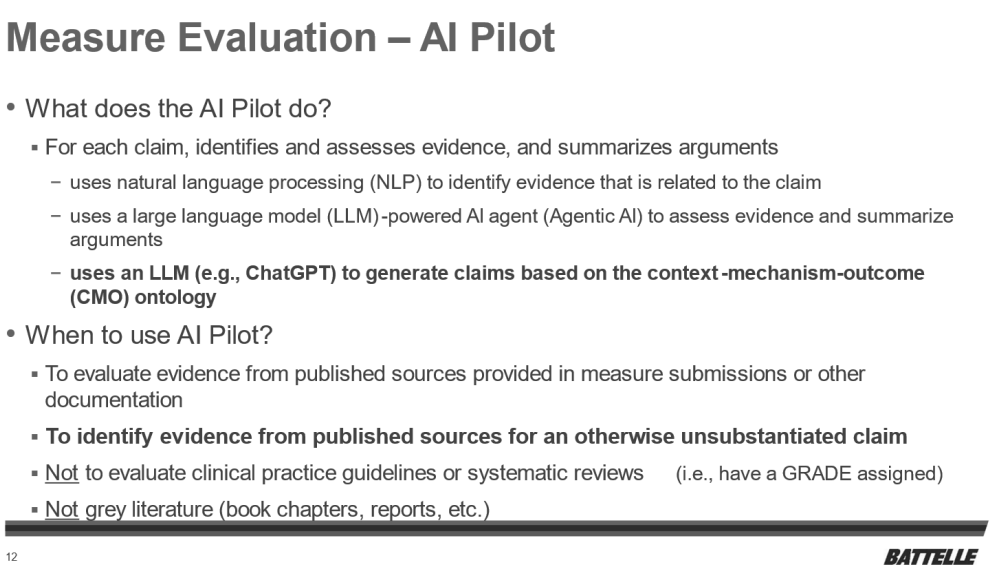
and we sort of make a claim that what's causing that association is quality, but the "mechanism claim," you need both an association and a mechanism claim. The mechanism claim is basically that we can actually describe what is responsible for that association in terms of these responses and the sources.

[SLIDE 11]



GEPPERT: Finally, this is kind of the structure of an assurance case. The point that I want to just raise here again is that this always is kind of residual risk in an assurance case. The idea is to make a decision with full knowledge of what the residual risks are, and then to be able to mitigate those residual risks and sort of develop a plan going forward.

[SLIDE 12]

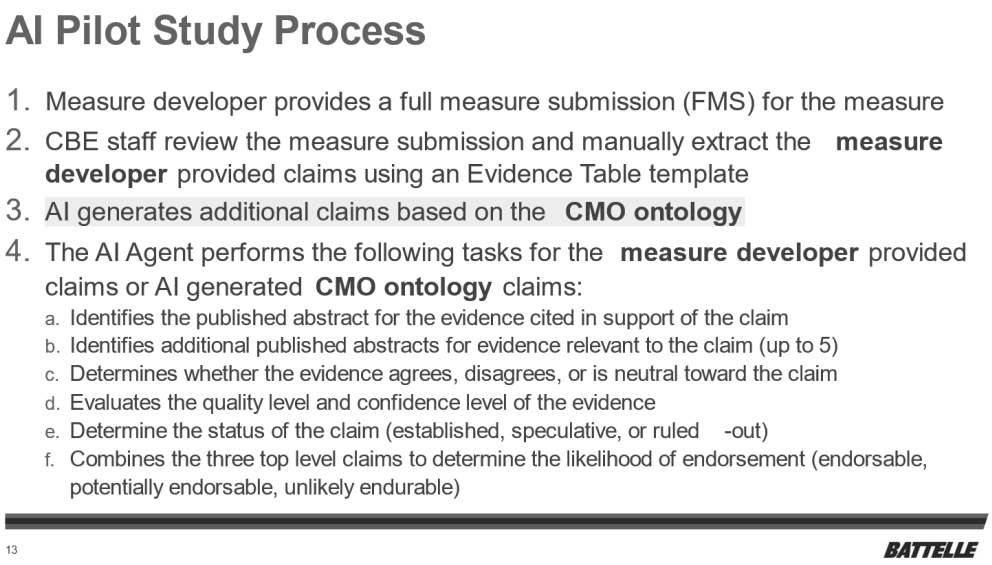


GEPPERT: So for the pilot our goal was really to use a process that was transparent and replicable. We didn't want to use any sort of technology or software or analysis that wasn't generally available, and there were two reasons for this. One is we wanted you to be able to replicate what we're doing, and so everything that we go through today, you are able to do it yourself if you have access to a generally trained LLM like ChatGBT or some other model. So we encourage you to do that, and to experiment and to follow the process that we're describing here to become familiar with this yourself.

The second reason for doing that in terms of the AI Pilot was that it sort of established a baseline for the performance of the generally trained models to understand what the limitations are, what the capabilities are, and then to inform in a data-driven way sort of where there might be some benefit in

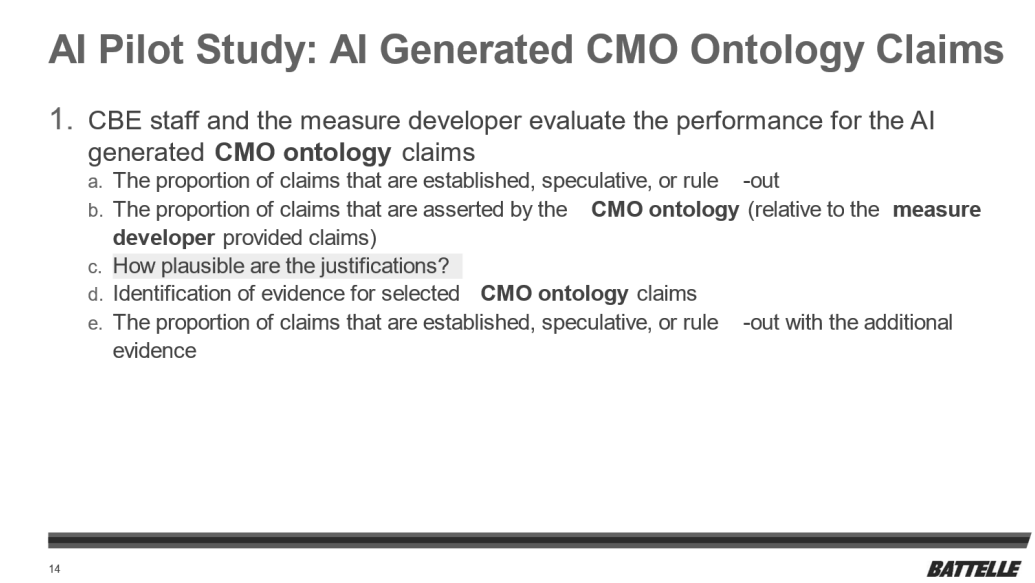
investing in more sophisticated models or methods or different approaches. So those were sort of the two aims of the AI Pilot.

[SLIDE 13]



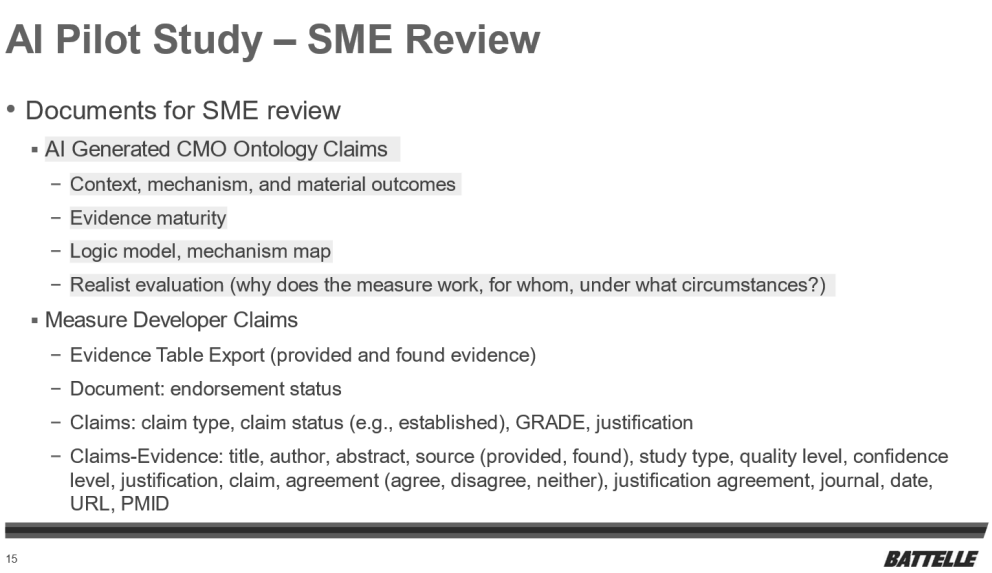
GEPPERT: So I won't go into the details of the AI Pilot, but basically what we're talking about is we have a measure submission. We have a measure description, and we use AI to generate claims about that measure using this claim mechanism outcome ontology.

[SLIDE 14]



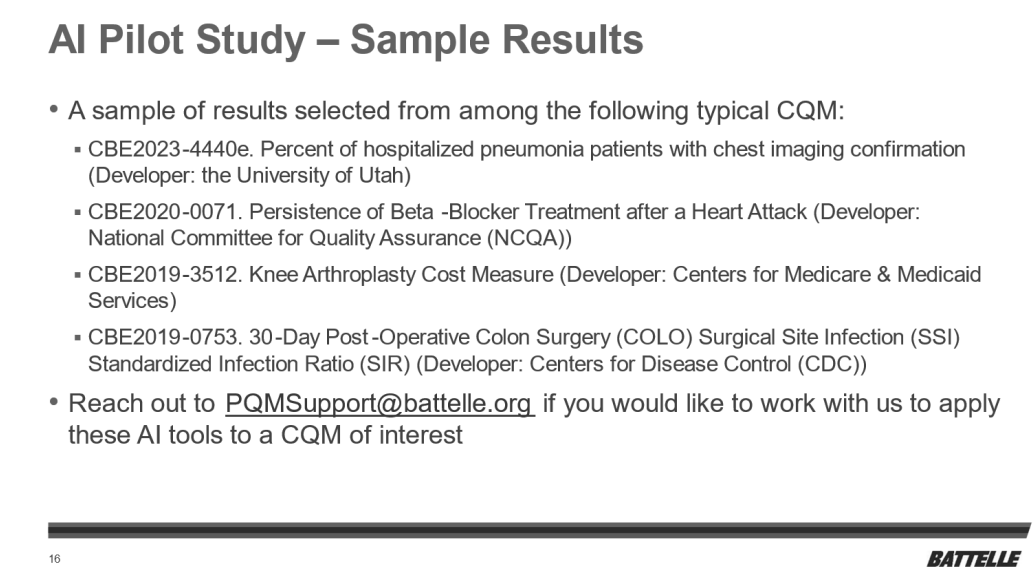
GEPPERT: And what we're really doing is looking at the results that the AI generates and judging them based on their plausibility and their logic from the perspective of a subject matter expert (SME) who knows about the material. And then if there are any claims, you know, that again are potentially speculative or arguably false, then those can be sort of followed up with further evidence generation.

[SLIDE 15]



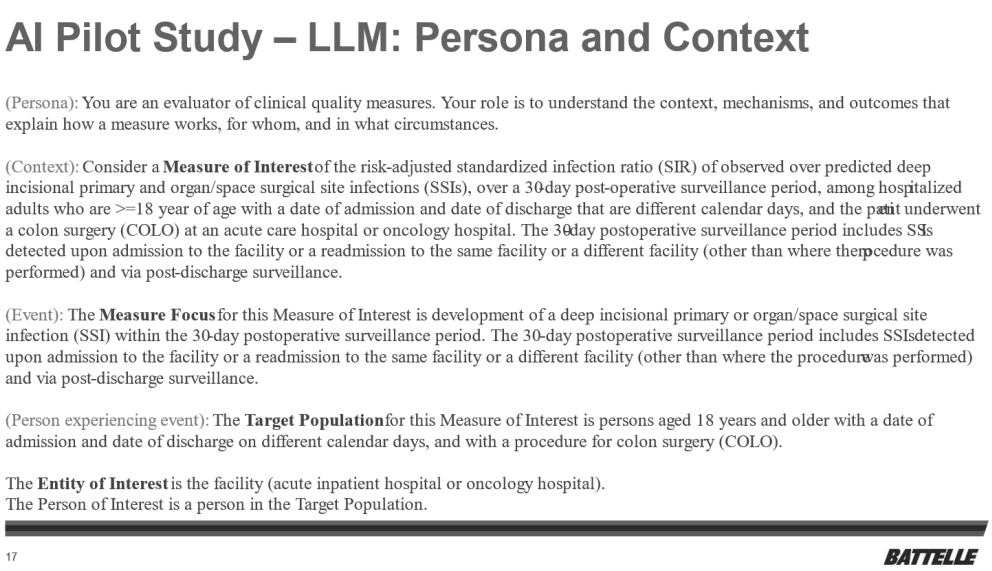
GEPPERT: And this is kind of the structure of what we're looking at, these contact mechanism and material outcome claims, evidence maturity claims, a logic model, and a mechanism map. And then finally, this realist evaluation. Why does the measure work, for whom, and under what circumstances?

[SLIDE 16]



GEPPERT: These are some of the measures we're going to be looking at. We've looked at several hundred measures as part of this pilot. We've picked a couple for some sample results, a couple of process measures, an outcome measure, and a cost measure, but we very much encourage you to reach out to us. If you'd like to learn more, if you'd like to work with us more on this, we'd be happy to do that. Help try these tools out and see how informative it is to your measure use. You don't have to be a measure developer in order to use these tools. You could be a measure user, or you could have a measure that's of interest to you, and these tools can help you understand how the measure is intended to work and what some of the challenges might be.

[SLIDE 17]



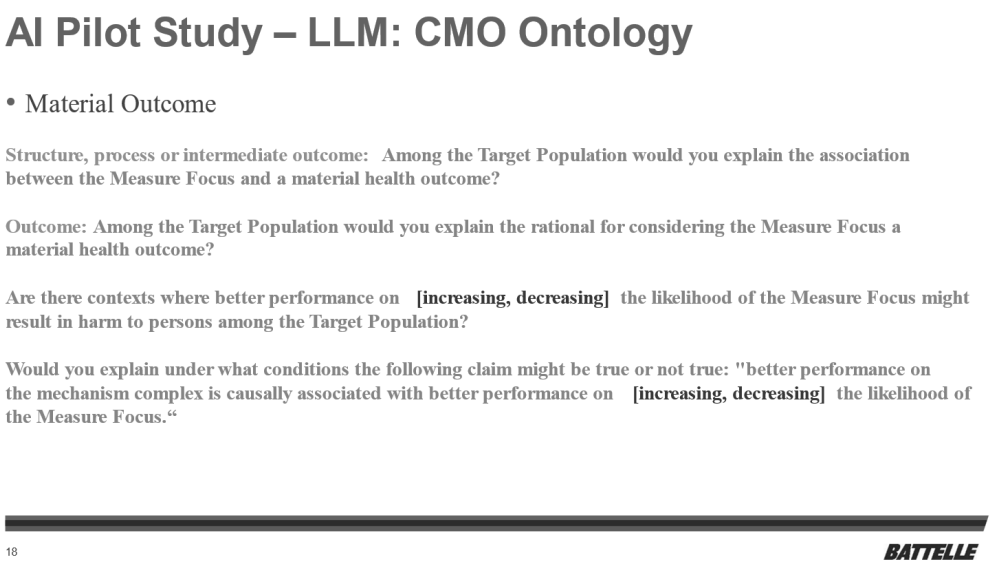
GEPPERT: Okay, now we're going to get more into some of the specifics. So when using a large language model like ChatGBT, there are certain ways that you can structure the questions, the prompts that you're asking ChatGBT in order to improve the responses that you receive. So some of that is here, sort of defining a persona and a context. Large language models (LLMs) are very context-sensitive, and so it's important that you sort of establish the context of your prompts very specifically.

What we have found works well is organizing the prompts sort of along these lines. So one is to establish who you are. You're an evaluator of clinical quality measures (CQMs), your willingness to understand the context mechanism and outcomes that explain how the measure works and through what circumstances. The context is you have a measure of interest and then you describe the measure of interest. So basically the

descriptions that are typically provided about measures. And then you want to define sort of the measure focus and the target population. What we have found works well here is to describe the measure focus as an "event," and so something that happened in a particular period of time. So, in this case, it's a surgical site infection (SSI) event that happens in a particular point in time.

Sometimes measures are *continuous*. They might be like an observed to expected ratio or cost measure. Even it's helpful to define it as an event as being what the person experiences, you know, cost that's more than what's expected, or an outcome that's worse than what's expected, but describing a measure focus as an event and then the target population is described as a person who's experiencing that event. So you define the measure focus, define the target population, and then the entity of interest. So in this case it's a facility, or it could be a clinician, or it could be a health plan. And then we don't really do anything with this person of interest, but it's just helpful to be able to make sure, because we do refer to persons sometimes to make sure that it's understood that the person is in the target population.

[SLIDE 18]



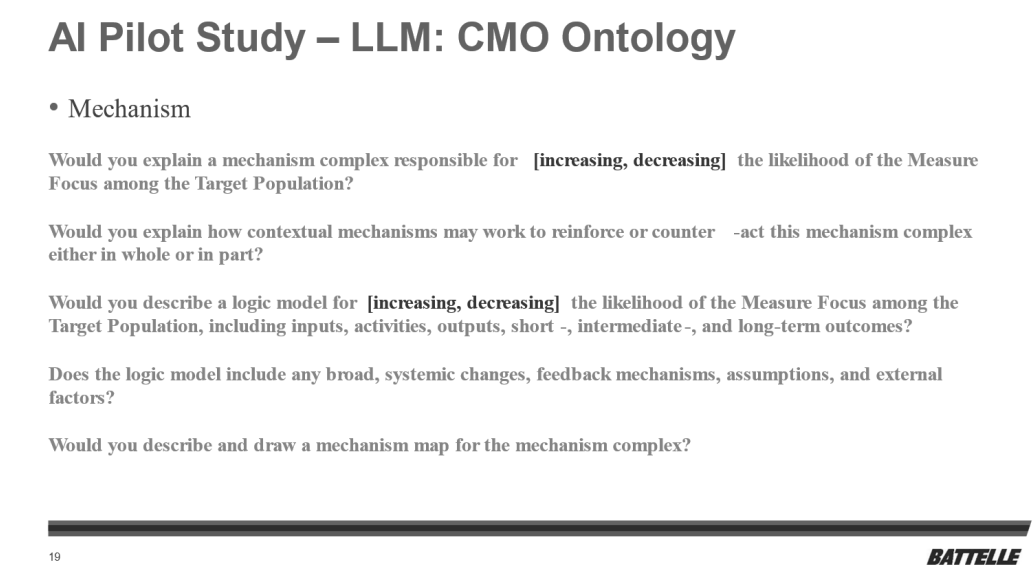
GEPPERT: So these are the prompts that we use. So the first part of the CMO ontology is the outcome. We have different prompts depending on whether the measure is a structure and process, or an intermediate outcome, or it's an outcome itself. So if it's a structure process or intermediate, we want to know among the target population which we've previously defined. We do explain the association between the measure focus and a material health outcome. So this is importance, right? We're looking for that association. If the measure is an outcome, like a readmission or a mortality, we want to ask among this target population what is the rationale for considering the measure focus on material health outcomes, and it provides sort of an explanation for why that is.

So that's the outcome part, and then we want to get to the harm part which are contexts where better performance on the year were either increasing the likelihood of the measure focus, or were decreasing, depending on whether it's adverse events or positive events. How that

might result in harm to persons among the target population, and so this would be a description of what are those potential harms that might occur based on better performance.

And then this is kind of the summary of it. "Would you explain it, or under what conditions the following claim might be true or not true." Better performance on this mechanism complex is causally associated with better performance on increasing or decreasing the likelihood of the measure focus. This is a good summary of the validity argument that's necessary in order to sort of establish that claim. So those are kind of the outcome-oriented.

[SLIDE 19]

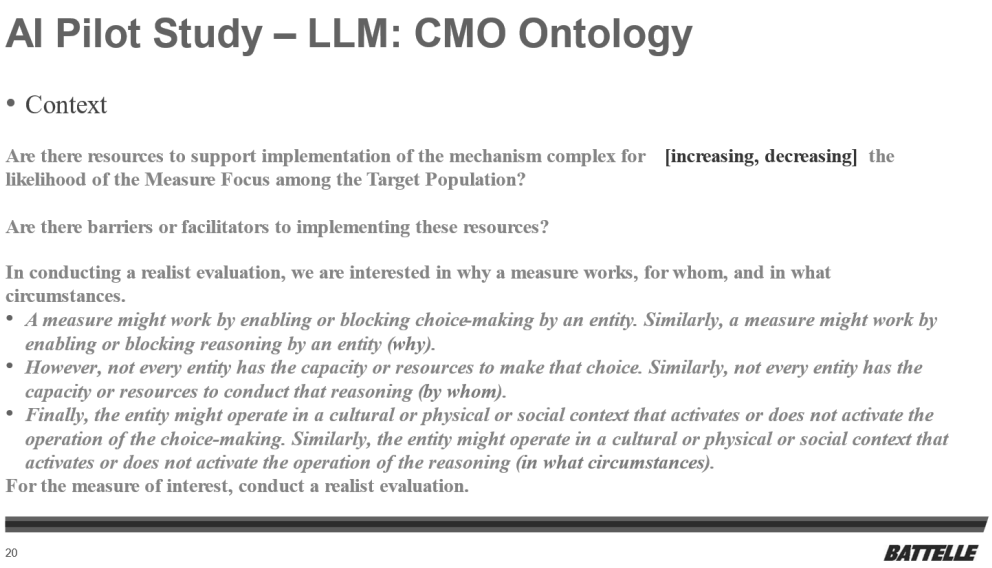


GEPPERT: So what is that mechanism complex? So the first prompt is "would you explain a mechanism complex responsible for either increasing or decreasing the likelihood of the measure focus amongst the target population?" That's kind of why we describe the measure focus as an "event," because we're interested in the likelihood of that event.

And then second, "would you explain how other contextual mechanisms may work to either reinforce or counteract this contextual, this mechanism complex?" So this is where we're just looking at the context and, you know, things that are either supportive or not supportive of a better performance on the measure. And then we ask it to build a logic model for either increasing or decreasing the likelihood of the measure focus. We describe what we want that logic model to contain — inputs, activities, outputs and short, intermediate, and long-term outcomes.

And then we ask whether there's any broad systemic changes, feedback mechanisms, assumptions or external factors that should be included in the logic model. And then finally, we want it to describe and draw a mechanism map for the mechanism complex. So mechanism complexes are complicated, but it's often sort of easier. The logic model is one way of describing it, and a mechanism map is another way of potentially describing it.

[SLIDE 20]



GEPPERT: So then finally we want to understand more about the usability of the measure, right? That's sort of what we're getting at with our context. "So, are there resources to support implementation of the mechanism complex?" We want to know about that. "Are there barriers or facilitators to implementing these resources?" And then here we're asking it to conduct what's known as a "realist evaluation." Why the

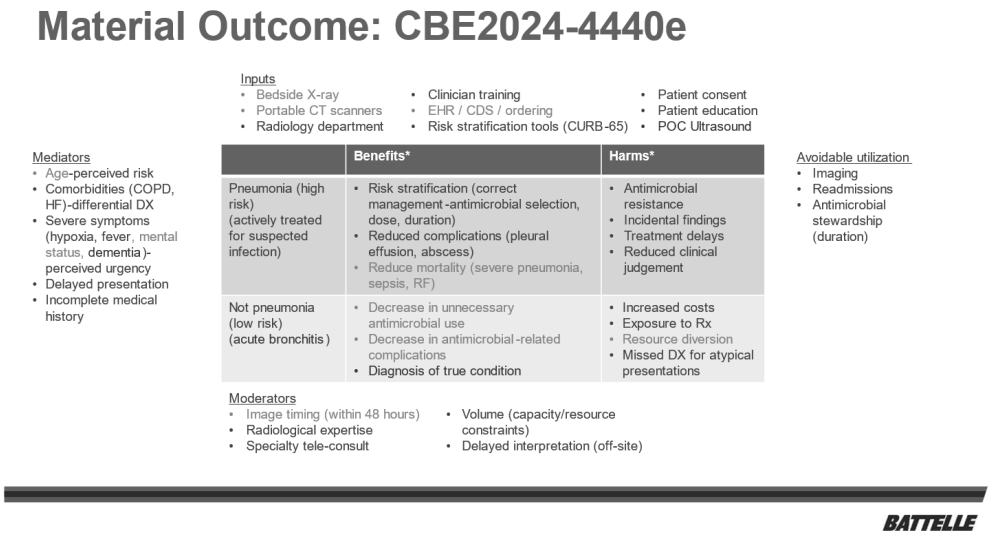
measure works, for whom, and in what circumstances. So there's kind of a specific way that we ask about this.

So the measure works in basically two different ways. Again, resources and response to those resources. So we're interested in knowing how does the measure either enable a choice or block a choice you don't want people to make? How does it do that? So we want a description of that. Similarly, it might either enable or block some type of reasoning that the entity might use. We want to know how it does that.

So that's sort of the "why" question. The "by whom" question is not every entity has the capacity or the resources to make the choice or conduct that reasoning. For whom might that be a problem? Finally, what are the circumstances? There are all sorts of ways of describing context. We talk about cultural, physical, or social contexts. There are different contextual models that one might use, but what we're interested in is how that context either activates or does not activate the operation of the choice-making. That's sort of the "realist" things that choice-making is triggered by this context, similarly around the reasoning and what circumstances.

So again, the goal of this is to say how the measure works, for whom does it work, and the circumstances in which it works. Okay, now we're going to look at some of the results.

[SLIDE 21]

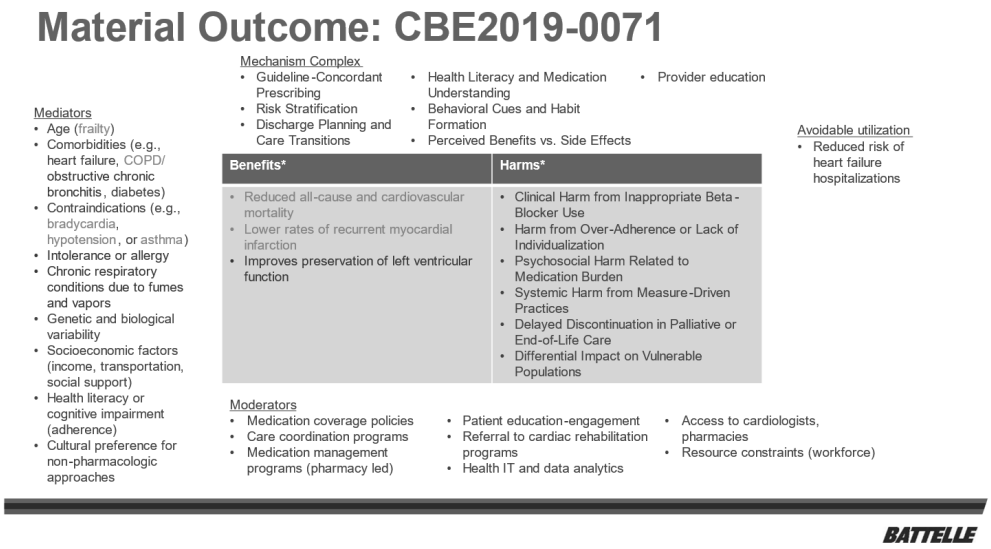


GEPPERT: So this is a summary of a lot of the CMO kind of output and put into a table just to make it a little bit easier to look at. So the point here is that this is all AI-generated, all of these results. So the AI will identify potential mediators that are relevant to this particular measure. It will describe the mechanism complex, which is kind of the top part. "What it is that the clinician or the entity needs to do in order to increase the likelihood of this particular measure? Is it focused on the diagnosis of pneumonia?" It will talk about potential harms associated with the measure, utilization, benefits and harms, and utilization that could potentially be avoided through better performance on the measure. It will talk about these contextual factors that either make the mechanism complex work better, or less better given the context.

And then it talks about these benefits and harms. "So what are the benefits to the patient by better performance on the measure, but also what are potential harms associated with better performance on the measure?" This is probably where our experience has been that large language models (LLMs) are particularly useful, in terms of articulating these harms which are otherwise a little bit challenging to identify.

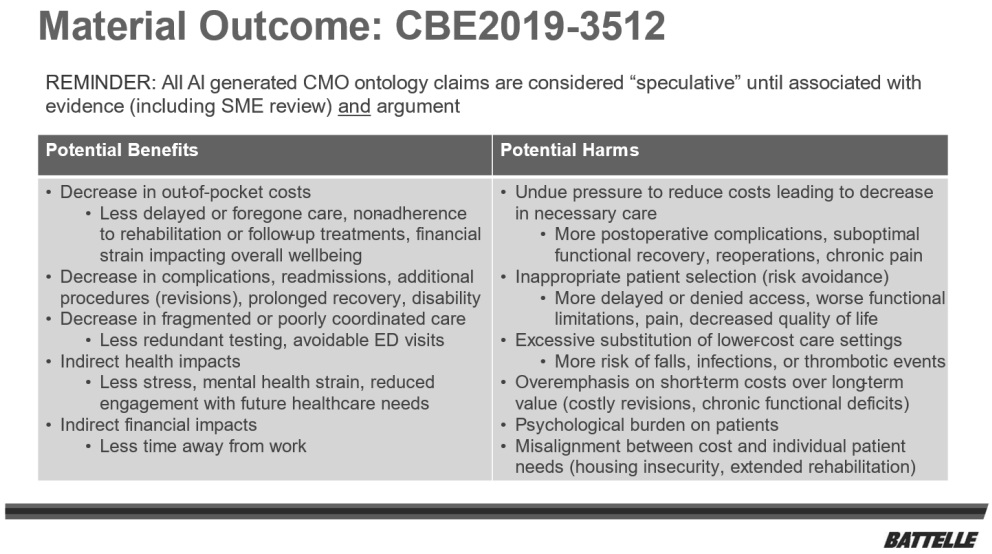
The other aspect about a diagnosis measure that is of interest is there are benefits and harms — not only for those that are appropriately diagnosed, but also those that are appropriately not diagnosed with pneumonia, but are appropriately diagnosed with something else. But then the big thing to remember here is that these claims generated by a large language model (LLM) according to that diagram are what we call "unsubstantiated claims." So these need to be reviewed by an SME for plausibility. And if they're either considered to be speculative or arguably false, then there needs to be sort of evidence generation that follows up with this. That's an important consideration.

[SLIDE 22]



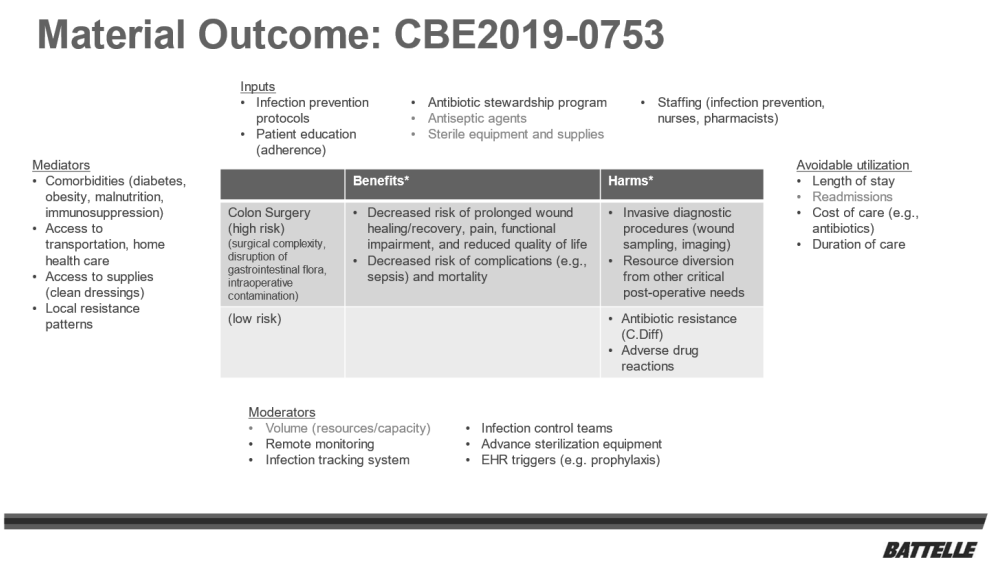
GEPPERT: So this is sort of another measure. This one is a process measure related to beta blocker prescriptions after an AMI. So again, it identifies all of the potential mediators, the patient factors that could influence this measure. It might have to be adjusted for—it identifies what the mechanism complex is, what the avoidable utilization is, the benefit of the measure, some other mediators that might impact the performance of that mechanism complex. And then some of the potential benefits and harms, you know, associated with the measure. So this would be sort of a process measure, and it gives a very interesting sort of definition of harms. Psychological harms it identifies and things like that nature.

[SLIDE 23]



GEPPERT: So this is a cost measure, and it talks again about the benefits and harms to the patient, decrease in out-of-pocket costs, decrease in complications, decrease in fragmented and poorly coordinated care, but then potential harms about potential underutilization of needed care. So it describes all of those things.

[SLIDE 24]



GEPPERT: So the purpose of this is not really to go into the results in any kind of *specificity*. What we're really trying to demonstrate is what's possible with this AI—all of the results that AI is capable of generating and that gives us sort of a flavor or a sense of that. This again is an outcome measure which contributes to these different things.

[SLIDE 25]

Mechanism: CBE2024-4440e Logic Model

Inputs (Resources-Means)	Activities (What the program does-Ways)	Outputs (Direct results of the activities)	Outcomes	Impact (Broad, systemic changes influenced by the quality program):
1. Staff and Expertise: - Radiologists, imaging technicians, and clinicians trained in pneumonia diagnosis and management. - Quality improvement (QI) teams and clinical leaders. 2. Infrastructure: - Imaging equipment (e.g., X-rays, CT scanners, portable units). - Electronic Health Record (EHR) systems with integrated clinical decision support (CDS) tools. 3. Financial Resources: - Funding for equipment, staff training, and operational support. - Grants or reimbursement incentives tied to pneumonia care measures.	1. Protocol Development: - Develop or refine protocols requiring timely chest imaging for suspected pneumonia cases. 2. Training and Education: - Conduct training sessions for clinicians on evidence-based pneumonia diagnosis, imaging indications, and antimicrobial stewardship. - Educate patients about the role of imaging in pneumonia care. 3. Technology Integration: - Deploy CDS tools in the EHR to prompt imaging orders based on clinical criteria. 4. Quality Improvement (QI) Initiatives: - Launch PDSA (Plan-Do-Study-Act) cycles to test and refine interventions aimed at increasing imaging adherence.	1. Protocols Implemented: - Standardized imaging protocols for pneumonia patients adopted across facilities. 2. Trained Staff: - Clinicians, radiologists, and QI teams trained in pneumonia care and imaging processes. 3. Functional Technology: - CDS tools operational in EHR systems, prompting timely imaging. 4. Engaged Patients: - Increased patient understanding of imaging's role in pneumonia care. 5. Data Availability: - Regular performance reports on imaging rates and adherence to protocols.	Short-term (Changes resulting from the outputs) 1. Increased adherence to imaging protocols for pneumonia diagnosis. 2. Improved clinician awareness and confidence in pneumonia care guidelines. 3. Higher rates of timely chest imaging among the Target Population. 4. Enhanced patient acceptance of imaging recommendations. Intermediate term (Effects observed as the program matures) 1. Clinical Process Improvements: - Reduced variability in imaging practices across clinicians and facilities. - More accurate pneumonia diagnoses due to consistent imaging use. 2. Resource Optimization: - Efficient use of imaging equipment and staff.	1. Diagnostic Standardization Across Facilities: - Implementation of uniform imaging protocols for pneumonia diagnosis within healthcare networks, improving consistency and reducing variability in care. 2. Integrated Care Models: - Enhanced coordination between primary care, emergency departments, radiology, and inpatient teams to streamline diagnostic workflows and improve efficiency. 3. Focus on Value-Based Care: - Shift toward reimbursement models that emphasize quality and outcomes (e.g., reduced mortality, fewer readmissions) rather than volume of services, aligning incentives with better adherence to the Measure Focus. 4. Expansion of Telehealth and Remote Diagnostics: - Increased use of tele-radiology and mobile imaging to improve access in underserved or rural areas. 5. Data-Driven Healthcare:

BATTELLE

GEPPERT: So that's sort of the outcome part of it, the benefits, the harms, the patient factors that are important. This is the kind of logic model that the AI is able to generate. Again, we've told it what we want in terms of these column headings, but all of the content is generated by AI. One thing I wanted to highlight here is that this particular measure is a process measure, and according to our CMO ontology the outcome part in a logic model for a process measure.

So outcome doesn't mean the outcome, you know, that the process measure is related to. The outcome for purposes of the logic model is the process measure, measure focus itself. So in this case increased adherence to the protocols for diagnosis. We have a link to some of the logic model guidance that PQM has been building to help folks develop and interpret these logic models.

[SLIDE 26]

Mechanism: CBE2019-0071 Logic Model

Inputs (Resources-Means)	Activities (What the program does-Ways)	Outputs (Direct results of the activities)	Outcomes	Impact (Broad, systemic changes influenced by the quality program):
• **Clinical Guidelines:** ACC/AHA recommendations for post-AMI care. • **Health IT Systems:** EHRs with clinical decision support (CDS), pharmacy databases, adherence monitoring tools. • **Healthcare Workforce:** Physicians, nurses, pharmacists, case managers, and care coordinators. • **Patient Resources:** Education materials, medication adherence apps, pill organizers.	• **Provider-Focused Activities:** ○ Training on guideline-based prescribing. ○ Integrating CDS tools into EHRs for beta-blocker reminders. ○ Regular performance feedback and audit reports. • **Patient-Focused Activities:** ○ Medication counseling at discharge and follow-up visits. ○ Providing adherence tools (pillboxes, mHealth apps). ○ Motivational interviewing to address barriers to adherence. • **System-Level Activities:** ○ Standardizing discharge protocols (e.g., Project RED, BOOST). ○ Care coordination between inpatient and outpatient providers.	• **Quantitative Outputs:** ○ Number of providers trained on beta-blocker guidelines. ○ Number of patients receiving discharge medication counseling. ○ Number of CDS alerts generated and acted upon. ○ Percentage of eligible patients enrolled in adherence programs. • **Qualitative Outputs:** ○ Improved provider knowledge and confidence in managing post-AMI medications. ○ Enhanced patient understanding of the importance of beta-blocker therapy.	Short-term (Changes resulting from the outputs): • **Increased Provider Awareness:** Greater adherence to guidelines for prescribing beta-blockers at discharge. • **Improved Patient Knowledge:** Patients understand the role of beta-blockers in preventing future cardiac events. • **Enhanced Medication Access:** Reduced financial barriers through insurance coverage and prescription assistance programs. • **Initial Adherence:** Higher rates of prescription fills within the first 30 days post-discharge. Intermediate term (Effects observed as the program matures) • **Sustained Medication Adherence:** Increased percentage of patients meeting the persistence	• **Transition to Value-Based Care:** ○ Movement from fee-for-service models to value-based payment systems (e.g., Medicare Advantage Star Ratings, ACOs) incentivizes medication adherence as a quality metric. ○ **Impact:** Aligns financial incentives with persistent beta-blocker use, encouraging health plans and providers to invest in adherence programs. • **Expansion of Health IT Infrastructure:** ○ Nationwide adoption of electronic health records (EHRs) with clinical decision support (CDS) tools, health information exchanges (HIEs), and integrated pharmacy data systems. ○ **Impact:** Facilitates real-time monitoring, adherence tracking, and

BATTELLE

GEPPERT: So then the activities part is more of the mechanism part. So these are the things that an entity needs to do in order to improve the likelihood of the measure focus. So there are provider-focused activities in the logic model. There are patient-focused activities, system-level activities, all of which are sort of necessary in order to achieve the outcome of interest.

[SLIDE 27]

Mechanism: CBE2019-3512 Logic Model

Inputs (Resources-Means)	Activities (What the program does-Ways)	Outputs (Direct results of the activities)	Outcomes	Impact (Broad, systemic changes influenced by the quality program):
• **Clinical Guidelines & Toolkits:** AAOS guidelines, ERAS protocols, BPCI/CJR resources. • **Clinical Staff:** Surgeons, anesthesiologists, nurses, physical therapists, case managers. • **Care Coordination Staff:** Care navigators, discharge planners, social workers. • **Health IT Systems:** Integrated EHRs, predictive risk tools, data tracking systems. • **Financial Support:** Value-based payment models (e.g., Bundled Payments), organizational leadership commitment. • **Patient Education Tools:** Shared decision aids, prehabilitation materials, discharge instructions.	• **Preoperative Optimization & Risk Stratification (30 Days Before Surgery):** ○ Comprehensive patient risk assessment (e.g., BMI, diabetes control, frailty screening). ○ Prehabilitation (prehab) involving physical therapy, nutrition optimization, smoking cessation. ○ Patient education and shared decision-making to align expectations. ○ Care planning for patients with complex needs (e.g., SDOH support). • **Standardized Surgical & Anesthesia Protocols (During Surgery):**	• **Increased Preoperative Risk Assessments & Prehab Completion:** % of patients receiving pre-surgical risk screening and prehab. • **Standardized Intraoperative Practices:** % adherence to ERAS protocols and multimodal pain management. • **Timely Discharge & Recovery Plans:** % of patients discharged home with clear instructions and support. • **Patient Engagement:** % of patients reporting understanding their care plan. • **Postoperative Follow-up:** % of patients receiving timely follow-up and navigation support.	Short-term (Changes resulting from the outputs): • **Fewer Preventable Complications:** Reduced rates of surgical site infections, venous thromboembolism (VTE), and other common post-surgical complications. • **Reduced Length of Stay (LOS):** More patients safely discharged on the day of or day after surgery. • **Improved Patient Activation:** Patients more informed and confident in managing their recovery. • **Improved Care Transitions:** More patients experience seamless handoffs from surgery to home rehabilitation. Intermediate term (Effects observed as the program matures) • **Lower Readmission Rates:** Fewer patients returning to the hospital for preventable issues. • **Reduced Post-Acute Facility Use:** More patients recovering safely at home rather than	• **Normalization of Value-Based Care**: Care coordination, ERAS pathways, and episode management become standard practice across orthopedic surgery. • **Shift Toward Multidisciplinary Surgical Teams**: Surgeons, anesthesiologists, physical therapists, and care navigators routinely co-manage surgical patients, increasing team-based care culture. • **Greater Emphasis on Patient Activation**: Systems increasingly invest in tools and processes to involve patients in their preoperative preparation and postoperative recovery. • **Expansion of Preoperative Optimization as Routine**: Prehabilitation and risk stratification protocols become a default part of surgical preparation for other procedures beyond knee arthroplasty. • **Health Equity Integration**: Organizations address social

BATTELLE

GEPPERT: I wanted to highlight that output column because it's sort of interesting to see. One of the challenges obviously is with defining these mechanisms as they can be complex. They can be challenging to measure and to quantify, but the output column gives some sort of quantifiable empirical things that one should see, if the mechanism complex is being adhered to. So if you think of like a structural measure, this would be sort of a useful column. "What should you be able to measure and quantify if, in fact, the entity is following the mechanism complex as they should, as the theory suggests they should?"

So I think this is something that we could be using more in cases where maybe we don't observe the mechanism directly, but we observe some of the consequences of the mechanism.

[SLIDE 28]

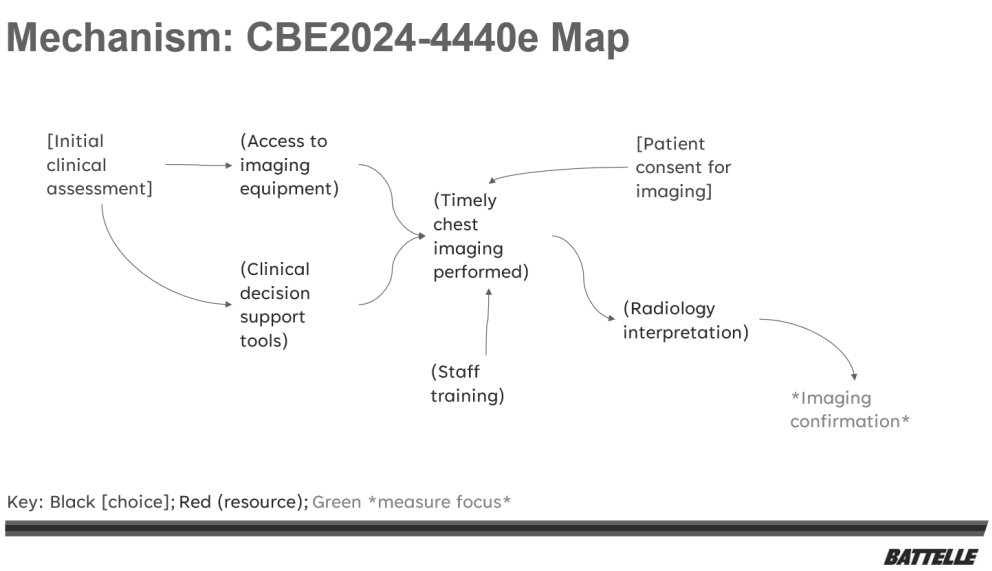
Mechanism: CBE2019-0753 Logic Model

Inputs (Resources-Means)	Activities (What the program does-Ways)	Outputs (Direct results of the activities)	Outcomes	Impact (Broad, systemic changes influenced by the quality program):
• **Human Resources:** ○ Infection preventionists, surgeons, nurses, and pharmacists. ○ Educators for staff and patient training. ○ Administrative support for program oversight. • **Financial Resources:** ○ Funding for training, surveillance systems, and procurement of supplies. • **Infrastructure and Tools:** ○ Sterile surgical equipment and supplies (antiseptics, gowns, drapes). ○ Technology (e.g., electronic health records [EHRs] with alerts, infection tracking tools). ○ Access to evidence-based guidelines (CDC, WHO).	• **Training and Education:** ○ Regular training sessions on SSI prevention for healthcare staff. ○ Patient education on preoperative preparation and post-discharge wound care. • **Preoperative Interventions:** ○ Risk stratification for high-risk patients. ○ Implementation of evidence-based antisepsis and bowel preparation protocols. ○ Timely administration of prophylactic antibiotics. • **Intraoperative Practices:** ○ Adherence to sterile surgical techniques. ○ Maintaining optimal operating room conditions (e.g., airflow, sterilization).	• **Short-Term Outputs:** ○ Number of staff trained in SSI prevention protocols. ○ Number of patients educated on wound care and follow-up. ○ Implementation of standardized preoperative, intraoperative, and postoperative protocols. ○ EHR-enabled reminders and documentation for key interventions. • **Intermediate Outputs:** ○ Increased adherence to infection prevention guidelines. ○ Improved communication and collaboration among multidisciplinary teams.	Short-term (Changes resulting from the outputs): • Improved compliance with preoperative antibiotic prophylaxis and sterile technique. • Early detection and intervention for postoperative wound issues. • Increased staff knowledge and confidence in SSI prevention. Intermediate term (Effects observed as the program matures) • Reduction in the rate of SSIs identified during initial hospitalization or post-discharge. • Improved patient satisfaction and trust in surgical care.	• **Cultural Shifts:** ○ Promoting a culture of safety and accountability in healthcare settings. ○ Encouraging a mindset of continuous quality improvement among staff. • **Institutionalization of Protocols:** ○ Embedding standardized SSI prevention protocols into routine clinical workflows. ○ Establishing infection prevention as a core component of surgical care. • **Health Equity Improvements:** ○ Addressing disparities in infection prevention resources and access to care. ○ Tailoring interventions to underserved populations. • **Policy and Incentive Structures:** ○ Advocating for alignment of hospital policies with

BATTELLE

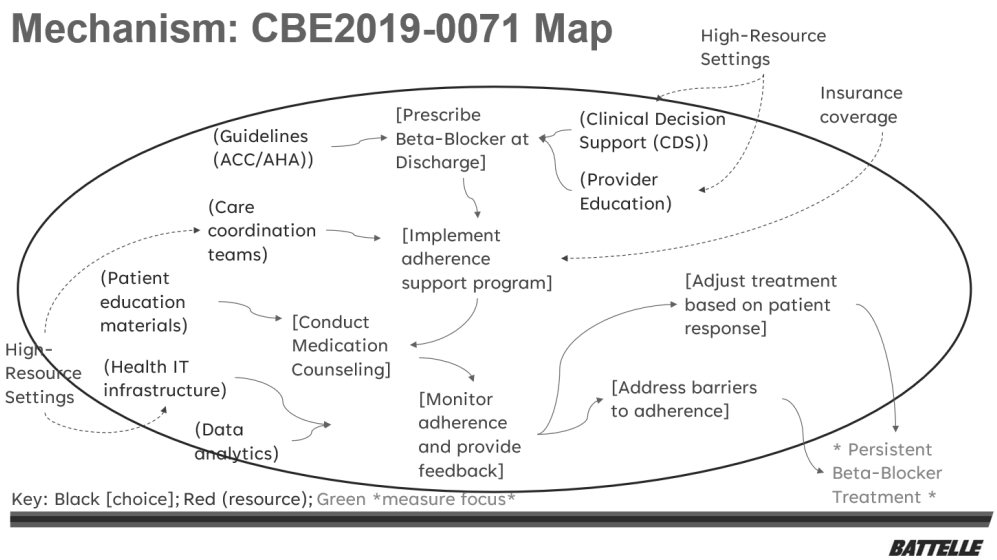
GEPPERT: Finally, really understanding what resources are necessary. Again, this gets to sort of the usability part. So we can define what it is that the entity should do in order to increase the likelihood of the measure focus, but are there necessary inputs that are challenging for certain entities and to define what those are, and to identify what the barriers are to those inputs, and to think about how those barriers could be mitigated.

[SLIDE 29]



GEPPERT: So this is an example of what a mechanism complex map looks like. So again, it consists of choices and resources, and so this is the one for the imaging-confirmed diagnosis. It starts with a *choice*. It generates some resources. Those resources beget other resources, or they leverage other resources. And then the end product of the mechanism complex is the measure focus.

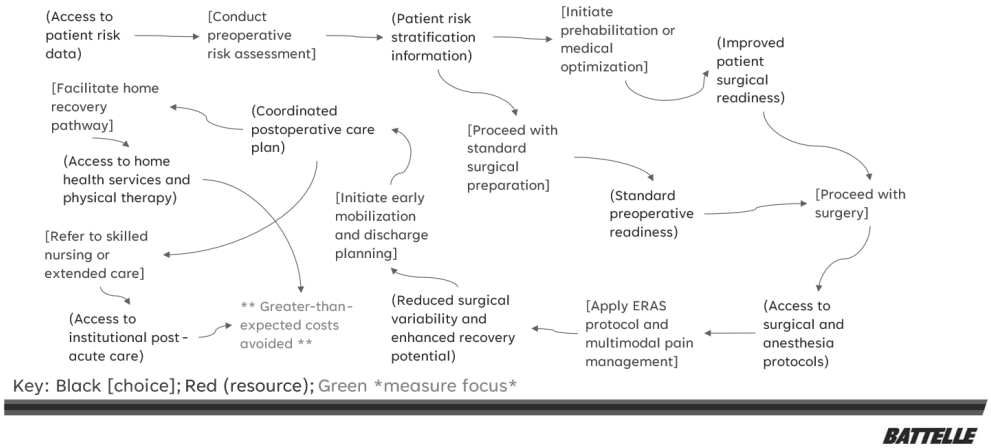
[SLIDE 30]



GEPPERT: This is another example, a little more complicated example. A process measure again where the circle is representing sort of, you know, what's within the entity and then what's outside the entity or some of the contextual factors that make a difference. And then the endpoint of the mechanism map is always sort of the measure focus for the measure.

[SLIDE 31]

Mechanism: CBE2019-3512 Map



GEPPERT: Here's another one for the cost measure, and the whole or the intention, you know, of sort of making these mechanism maps more explicit is a lot of your discussion when we have measurement meetings is around this and trying to articulate this. So we want to be able to make this more explicit in order to inform our conversations around what it is that's really necessary in order to again achieve the measure focus, which in this case is greater than expected costs.

[SLIDE 32]

Mechanism Complex Complexity Assessment

Complexity Domain	Cochrane Intervention Complexity Assessment Tool for Systematic Reviews (iCAT_SR)
Number of Components	**High Complexity** : The mechanism complex consists of multiple interdependent components spanning the preoperative, perioperative, and postoperative phases. These include risk assessment, prehabilitation, ERAS protocols, multimodal pain management, discharge planning, and post -acute care coordination.
Degree of Interaction between Components	**High Complexity** : The components interact dynamically —preoperative optimization affects surgical outcomes, which in turn influences postoperative recovery and discharge choices. Failures in one component (e.g., inadequate prehab) can cascade into downstream complications and increased costs.
Number and Variability of Outcomes	**Moderate Complexity**: While the primary focus is on cost reduction (Measure Focus), the mechanism complex also affects clinical outcomes (e.g., complications, readmissions, functional recovery) and patient experience (e.g., pain management, satisfaction).
Degree of Flexibility or Tailoring Allowed	**High Complexity** : The mechanism complex requires tailoring based on patient characteristics (e.g., age, comorbidities, functional status) and local resource availability (e.g., home health services in rural areas). Clinicians adapt pathways based on individual patient needs.
Context Dependency	**High Complexity** : Effectiveness is highly context -sensitive. Urban hospitals with integrated care teams and home health support may see cost reductions, whereas rural settings with post -acute care limitations may struggle to avoid SNF stays and higher costs.
Nature of the Behaviors Required by Those Delivering the Intervention	**High Complexity** : Successful implementation requires multidisciplinary teamwork across surgeons, anesthesiologists, physical therapists, case managers, and home health providers. It demands behavior change toward standardized protocols while allowing flexibility for patient -specific needs.
Nature of the Behaviors Required by Those Receiving the Intervention	**Moderate Complexity**: Patients are expected to engage in prehabilitation, adhere to recovery plans, and manage their rehabilitation at home. Older adults or patients with social barriers (e.g., housing insecurity, low literacy) may struggle with home recovery expectations, complicating adherence.
Potential for Spillover Effects	**Moderate Complexity**: Improvements in preoperative optimization and ERAS protocols can spill over into other surgical procedures, enhancing overall perioperative care standards. However, cost containment pressures can also spill over, leading to undercare in other patient populations.

Source: https://methods.cochrane.org/sites/methods.cochrane.org/files/uploads/icat_sr_additional_file_4_2016_12_27.pdf

BATTELLE

GEPPERT: So one thing we're exploring around these mechanisms is to assess their complexity, and there is a Cochrane tool that has these different dimensions to it that one can use to assess the complexity of a mechanism complex. The AI is capable of applying this tool to the mechanism complex, and it is described in assessing its complexity. Our thought is that trying to do this will be informative.

There may be some measures where the mechanism complex is highly complicated, other measures where it's not as complicated. That should inform our discussion, because it could inform how likely it is that the measure will improve, or the burden associated with that improvement. So this is an example of how AI can help us do things that would be very challenging to do sort of without being able to leverage the AI.

[SLIDE 33]

Context: CBE2024-4440e Realist Evaluation

Why Does the Measure Work?	For Whom Does the Measure Work?	In What Circumstances Does the Measure Work?
- The measure works by enabling or blocking choice-making related to whether the entity (facility) performs chest imaging for pneumonia patients. - It incentivizes adherence to evidence-based diagnostic protocols, prompting facilities to prioritize timely imaging as part of clinical workflows. - Choices include: - Ordering imaging based on clinical presentation. - Allocating resources to ensure imaging availability. - Incorporating imaging adherence into performance metrics. - The measure works by enabling or blocking reasoning regarding the appropriateness of chest imaging for pneumonia patients. - Reasoning processes include: - Assessing clinical presentation and determining the likelihood of pneumonia. - Evaluating risks and benefits of imaging for individual patients. - Balancing resource utilization against diagnostic accuracy.	- Not all entities have the capacity to make this choice due to resource constraints: - Well-Resourced Facilities: Large urban hospitals with advanced imaging technology and sufficient staffing are more likely to make this choice. - Resource-Limited Facilities: Rural or underfunded hospitals may lack imaging equipment or trained radiologists, limiting their ability to choose imaging as a diagnostic tool. - Capacity also depends on: - Staff training and clinical expertise. - Decision-making autonomy (e.g., influenced by institutional policies or payer constraints). - Not all entities have the resources or capacity to conduct this reasoning effectively: - Clinician Expertise: Facilities with experienced providers may perform more nuanced reasoning, leading to appropriate imaging decisions. - Technology Support: Entities equipped with clinical decision support (CDS) tools in electronic health records (EHRs) can enhance reasoning by automating guideline adherence and risk stratification. - Administrative Support: Hospitals with robust quality improvement systems may use performance data to refine reasoning processes over time. - Barriers to Reasoning: - High patient volumes or time pressures may limit the ability of clinicians to engage in careful reasoning. - Limited access to data or evidence-based guidelines can hinder informed decision-making.	- The cultural, physical, or social contexts that activate or inhibit choice-making include: - Cultural Context: A culture of defensive medicine may encourage overuse of imaging, while a culture of evidence-based practice ensures its appropriate use. - Physical Context: Geographic isolation or lack of infrastructure may prevent timely imaging, even if clinically indicated. - Social Context: Public health policies and payer systems that incentivize imaging adherence can activate the choice-making process, while lack of reimbursement or systemic inequities may inhibit it. - Reasoning is activated or inhibited by contextual factors: - Cultural Context: A culture emphasizing clinician autonomy may encourage reasoning based on experience, while strict adherence to protocols may discourage individualized reasoning. - Physical Context: Availability of imaging technology and rapid turnaround times can influence whether reasoning leads to imaging orders. - Social Context: Social norms or patient expectations may shape reasoning. For example: - Patients demanding imaging may pressure clinicians to comply. - Disparities in access to care or bias may result in differential reasoning for patients from underserved populations.

BATTELLE

GEPPERT: We won't spend a lot of time on this. These are realist evaluations. Again, we're focused on choices and reasoning, how does the measure work, by influencing choices, influencing reasoning, what kind of entities would be or persons would be challenged with performing those choices or reasoning. What context is either enabling or is a barrier.

[SLIDE 34]

Context: CBE2019-0071 Realist Evaluation

Why Does the Measure Work?	For Whom Does the Measure Work?	In What Circumstances Does the Measure Work?
• **How the Measure Works (Why):** The measure works by **enabling or blocking choice-making** by health plans, healthcare systems, and providers. This includes decisions related to: ○ **Prescribing behavior:** Whether to initiate beta-blocker therapy at discharge. ○ **Adherence support:** Whether to implement programs that encourage patient persistence with medication. ○ **Resource allocation:** Whether to invest in health IT, care coordination, and pharmacy services to improve adherence rates. • **Key Dynamics:** ○ **Enablers:** Clinical guidelines (e.g., ACC/AHA recommendations), performance incentives (e.g., HEDIS measures), and decision support systems. ○ **Blockers:** Lack of provider awareness, competing clinical priorities, and absence of financial incentives for adherence-focused interventions. • **How the Measure Works (Why):** The measure also works by influencing the **reasoning processes** of entities (health plans, providers): ○ **Clinical reasoning:** Evaluating the risk-benefit profile of beta-blocker therapy for each patient. ○ **Organizational reasoning:** Assessing the importance of medication adherence in quality improvement strategies. ○ **Policy reasoning:** Determining how performance on this measure impacts reimbursement, accreditation, or public reporting.	• **Large Integrated Health Systems:** ○ **Capacity:** Robust health IT infrastructure, dedicated QI teams, and strong care coordination models. ○ **Resources:** Access to comprehensive patient data, pharmacy integration, and financial resources to support adherence programs. • **Health Plans with Value-Based Care Models:** ○ **Capacity:** Sophisticated data analytics capabilities to monitor adherence and implement targeted interventions. ○ **Resources:** Incentive structures tied to quality measures (e.g., Medicare Advantage Star Ratings) that promote persistent medication use. • **Clinicians in Academic Medical Centers:** ○ **Capacity:** High levels of guideline awareness, access to continuing education, and multidisciplinary care teams. ○ **Resources:** Embedded clinical decision support tools and resources for medication counseling. • **Small or Resource-Limited Community Hospitals:** ○ **Capacity Constraints:** Limited health-IT capabilities, fewer staff dedicated to QI, and fragmented care transitions. ○ **Resource Gaps:** Lack of integrated pharmacy services and insufficient funding for adherence interventions.	• **Supportive Policy Environments:** ○ **Example:** Health systems operating under value-based payment models that reward medication adherence. ○ **Effect:** Strong alignment between financial incentives and adherence-improvement efforts. • **High-Functioning Care Transitions:** ○ **Example:** Hospitals with robust discharge planning, medication reconciliation, and follow-up care processes. ○ **Effect:** Increased likelihood that patients will receive, understand, and persist with beta-blocker therapy. • **Patient-Centered Cultures:** ○ **Example:** Organizations that prioritize shared decision-making, patient engagement, and culturally competent care. ○ **Effect:** Enhanced patient understanding, motivation, and adherence to prescribed therapies. • **Fragmented Healthcare Systems:** ○ **Example:** Lack of coordination between inpatient and outpatient providers, leading to gaps in medication management. ○ **Effect:** Increased risk of non-adherence due to poor follow-up and lack of continuity in care.

BATTELLE

GEPPERT: So these are just some different examples, but again the intention here is to inform our discussions around the measures to understand how it works and the environments in which it could be challenged to work to really inform the measure use. It could be that the measure works great in some situations and not so great in others, and maybe we ought to be going back to focusing our measurement resources where they can be the most impactful, you know, on areas where the measure works best.

[SLIDE 35]

Context: CBE2019-3512 Realist Evaluation

Why Does the Measure Work?	For Whom Does the Measure Work?	In What Circumstances Does the Measure Work?
Why: The Measure Works by Enabling or Blocking Choice-Making The measure influences clinician behavior by making episode costs visible and holding clinicians accountable for the financial efficiency of knee arthroplasty care. • It enables choice-making by highlighting cost drivers across the preoperative, perioperative, and post-acute care phases. • It blocks certain choices by incentivizing the reduction of unnecessary services and complications, discouraging overuse of post-acute facilities or prolonged inpatient stays. **Why: The Measure Works by Enabling or Blocking Reasoning** The measure promotes reasoning by prompting clinicians to evaluate their practice patterns and consider how preoperative preparation, surgical technique, and discharge pathways affect both cost and quality. • Enables reasoning by providing comparative cost performance data and stimulating reflection on variation across patients and providers. • Blocks reasoning when the cost signal is ambiguous or overly punitive, leading to defensive behavior (e.g., avoiding high-risk patients).	**For Whom: Capacity and Resources for Choice-Making** • Works well for: ○ Large, integrated health systems with care coordination capacity, data analytics infrastructure, and standardized surgical pathways. ○ Clinicians with access to home health, outpatient physical therapy, and prehabilitation programs. ○ Providers experienced with bundled payment models, who have support from administrative teams to optimize episode costs. • Works less well for: ○ Small practices and rural hospitals with limited administrative capacity, fewer care coordination resources, and gaps in home health or outpatient rehab. ○ Clinicians in fragmented systems where pre- and post-acute care decisions are made by different, unaligned entities. ○ Surgeons without data feedback or comparative benchmarks may lack the information necessary to make cost-conscious choices. **For Whom: Capacity and Resources for Reasoning** • Works well for:	**In What Circumstances: Context That Activates or Blocks Choice-Making** • Activating Contexts: ○ Bundled payment models (e.g., BPCI, CJR) align financial incentives with the measure, amplifying the choice-making mechanism. ○ Health systems with EHR-integrated pathways and care coordinators facilitating prehabilitation and discharge planning. ○ Professional culture emphasizing value-based care—clinicians view cost-efficiency as part of delivering high-quality care. • Blocking Contexts: ○ Fee-for-service environments encourage volume-driven care, weakening the choice-making mechanism. ○ Social determinants of health (SDOH) challenges (e.g., transportation barriers, housing instability) limit viable discharge choices, narrowing clinicians' options. ○ Regions with limited home health services or post-acute care deserts constrain discharge planning, reducing the effectiveness of cost-conscious decision-making. **In What Circumstances: Context That Activates or Blocks Reasoning** • Activating Contexts: ○ Benchmarking against peers contextualizes cost data, enabling interpretation beyond individual patient variability.

BATTELLE

GEPPERT: This one focuses on some context. Obviously, if you're in a region that has limited access to community resources, particular measures are going to be much more challenging than other measures.

[SLIDE 36]

Context: CBE2019-0753 Realist Evaluation

Why Does the Measure Work?	For Whom Does the Measure Work?	In What Circumstances Does the Measure Work?
• The measure works by **enabling or blocking choice-making** through its emphasis on: ○ **Incentives:** External motivators, such as public reporting and financial penalties, push entities to prioritize SSI reduction. ○ **Defined Options:** The measure outlines specific intervention strategies (e.g., preoperative prophylaxis, sterile techniques), guiding entities in their infection prevention efforts. • The measure works by **enabling or blocking reasoning** through: ○ **Access to Evidence:** Guidelines and best practices provide entities with the knowledge to reason through SSI prevention strategies. ○ **Data Feedback:** Regular surveillance and reporting (e.g., through NHSN) allow entities to analyze trends and identify gaps in performance. ○ **Problem-Solving Frameworks:** Tools like root cause analysis encourage entities to reason about failures and develop corrective actions.	• The measure works for entities that have: ○ **Organizational Capacity:** Adequate staffing levels, including infection preventionists, trained clinicians, and support staff. ○ **Financial Resources:** Funding for training, equipment (e.g., advanced sterilization tools), and technology (e.g., electronic health records). ○ **Leadership Support:** Engagement from hospital leadership to allocate resources and enforce compliance. • **For whom the measure does not work:** ○ Entities lacking financial or human resources may struggle to implement the required practices. ○ Smaller facilities with limited infrastructure, such as rural or community hospitals, may face barriers in adopting the choice-making framework. • The measure works for entities with: ○ **Analytical Capacity:** Access to trained infection preventionists or quality improvement experts who can interpret data. ○ **Data Infrastructure:** Advanced EHR systems or surveillance tools to collect, analyze, and report SSI data. ○ **Commitment to Quality Improvement:** Organizations that embrace continuous learning and adaptive practices. • **For whom the measure does not work:**	• **Cultural Context:** ○ A culture of safety and accountability encourages proactive decision-making. ○ Conversely, a punitive or blame-oriented culture may discourage honest reporting and learning from SSIs. • **Physical Context:** ○ Entities in well-resourced urban centers may find it easier to implement the necessary interventions. ○ Geographic isolation or resource limitations in rural settings can inhibit successful choice-making. • **Social Context:** ○ Collaborative environments, where multidisciplinary teams work together, enhance choice-making. ○ Fragmented systems, with poor communication between surgeons, nurses, and infection control staff, may block effective choices. • **Cultural Context:** ○ An open, learning-oriented culture fosters critical reasoning and adaptation. ○ A rigid or hierarchical culture may limit staff input and creative problem-solving. • **Physical Context:** ○ Entities with centralized data systems and access to real-time feedback are better positioned to reason effectively.

BATTELLE

GEPPERT: This is just about what resources are supporting the reasoning or the choice-making.

[SLIDE 37]

Building a Validation Roadmap

Validity Claim (A: entity response to measure; B: measure focus)	Association studies (A clinical study for the claim that A is a cause of B repeatedly measures the values of a set of measured variables that includes the variables A and B)	Mechanism studies (A mechanistic study for the claim that A is a cause of B is a study which provides evidence of structure or features of the mechanism (M1) by which A is hypothesized to cause B (M2))
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 I Importance Table (Performance score by decile): • Observed association between entity (A) and GTE costs (B) Known Groups: • Observed association between group (A) and GTE costs (B) Related Process or Outcome: • Fewer complications, unnecessary services, optimized recovery, patient activation	LEVEL 3 (arguable true) Effectiveness of the mechanism complex • Association between elements of the mechanism complex and the measure focus Reinforcing Contextual Mechanisms: • Integrated health systems and care continuity Counteracting Contextual Mechanisms: • Fragmented Care Systems and Weak Care Transitions • Resource Constraints in Smaller or Rural Settings
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 (arguable false) Reliability Table (reliability by volume decile): • Proportion of entity-level variation explained by uncertainty and chance Confounders C associated with both A and B: • Proportion of entity-level variation explained by risk-adjustment (comorbidities, frailty, physical) • Social Determinants of Health (SDOH) Challenges (low income, rural)	LEVEL 4 Experimental or quasi-experimental studies establish that implementing the mechanism complex reduces greater-than-expected costs. LEVEL 5 The explicit mechanism complex is widely recognized as a best practice.

BATTELLE

GEPPERT: So I mentioned that, you know, one of the things we asked the model is to evaluate the claim that better performance on the measure is associated with better performance on the measure focus. So that sort of speaks to validity. So what we're able to do is take the LLM output and sort of build this table, which is our evidence maturity table. So we have what's called "Level 1, Level 2, Level 3, and Level 4" evidence.

Basically, the higher the level, the stronger the causality argument, and typically we're in this Level 1 area where we're looking at different associations with different related measures. We sometimes get down into this Level 2 area where we're sort of starting to rule out other potential explanations. And then what we're trying to get to is this kind of upper right, this Level 3 area where we're starting to consider more of the mechanisms responsible for those associations meant ultimately, you know, into a Level 4 type of situation where we're actually looking at some experimental or quasi-experimental results.

[SLIDE 38]

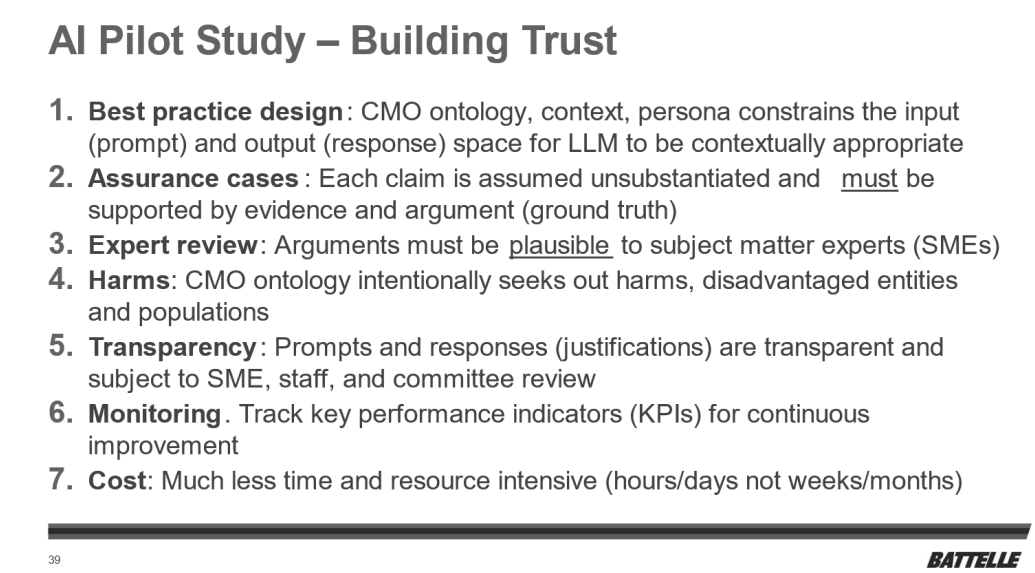
Building a Validation Roadmap

Validity Claim	Association studies (A clinical study for the claim that A is a cause of B repeatedly measures the values of a set of measured variables that includes the variables A and B)	Mechanism studies (A mechanistic study for the claim that A is a cause of B is a study which provides evidence of structure or features of the mechanism (M1) by which A is hypothesized to cause B (M2))
Causal claim (A is a cause of B)	(Prone to bias)	(Prone to complexity)
Specific mechanism hypothesis (posit features of such a mechanism complex)		Adoption/implementation (fidelity) of the mechanism complex: • Preoperative Optimization and Risk Stratification • Standardized Surgical and Anesthesia Protocols • Proactive Postoperative Care and Discharge Planning • Longitudinal Care Coordination and Case Management

BATTELLE

GEPPERT: That's just features of the mechanism.

[SLIDE 39]



GEPPERT: So I just want to end by talking a little bit about trust. So, you know, we're generating these results. "How do we know that we can rely on the results, that they're trustworthy?" So we think of that in different ways. So one is by *design*. So we use ontologies. We use assurance cases. We use context persona. All of these things work to constrain the input and output space for the large language model (LLM) to be contextually appropriate.

Our experience has been that largely that has been successful, that the responses are in fact contextually appropriate. Sometimes they might not be as detailed as you might like, and there are ways of going sort of deeper, but by and large we've seen very few instances of responses that just seemed contextually inappropriate.

Again, we're using assurance cases. So we're assuming that claims are unsubstantiated, and they're only substantiated if they're tied to a piece of evidence, as we described earlier in the argument. We always engage in expert review. All of these arguments must be considered *plausible*. We're specifically seeking out harms, and so we're looking at harms to the person, or certain contextual situations that might put people at a disadvantage. We're intentionally looking for those things. And then again, everything needs to be *transparent*. The prompts and the responses, all of which are subject to review.

Finally, we want to be able to track performance of this over time, and we'll look at an example of that. But then the ultimate sort of potential benefit is making measure development more *accessible.* Measure development can be very time and resource-intensive, and so one of the potential advantages of some of these AI tools is it just makes measure development more accessible to a broader population of folks that are interested in measurement, folks that want to develop measures in the community but maybe are less experienced. That's really the hope is to be able to enable that.

[SLIDE 40]

AI Pilot Study – System Safety

Comparison of the System-Theoretic Process Analysis (STPA) and Context-Mechanism-Outcome (CMO)

Phase	STPA	CMO
Define the purpose of the analysis	1) Identify loses 2) Identify system-level hazards 3) Identify system-level constraints 4) Refine hazards	1) A material outcome 2) Context-mechanism (worse) 3) Context-mechanism (better)
Model the control structure	1) Controller a) control algorithm b) process model 2) Controlled process	Quality Improvement Action Model 1) Agent: goal, authority, accountability a) feed-forward (goal-driven) b) feed-back (even-driven) 2) Action: perform-perceive
Identify unsafe control actions	1) Unsafe control action a) Not providing the control action leads to a hazard b) Providing the control action leads to a hazard c) Providing a potentially safe control action but too early, too late, or in the wrong order d) Control action lasts too long or is stopped too soon (for continuous control actions, not discrete ones) (source, type, action, context, hazard)	1) Context a) individual b) interpersonal c) institutional d) infrastructural (social, etc.) 2) CMO configuration (triggers)
Identify loss scenarios	1) Unsafe controller behavior 2) Inadequate control algorithm 3) Unsafe controller input 4) Inadequate process model	1) Wrong goal 2) Wrong execution/FF 3) Wrong perform-perceive 4) Wrong evaluation/FB

Source: MIT Partnership for Systems Approaches to Safety and Security (PSASS)

40 BATTELLE

GEPPERT: So I just want to give you a flavor for where we're going into kind of the long-term with a lot of this. So we really view quality measurement as a *system,* and the use of quality measurement as a system's issue, and so we want to make sure that the system works. So we want to be able to leverage frameworks that have been developed to ensure that systems are working as you intend them to work.

So one of these methodologies to ensure that the system works as intended is this STPA, System-Theoretic Process Analysis. Basically, the idea is that we want to understand how it is that the folks performing well — how do they do that? The folks that are performing less well, you know, how they do that. "What exactly needs to happen in order to transform a worse-performing entity into a better-performing entity?"

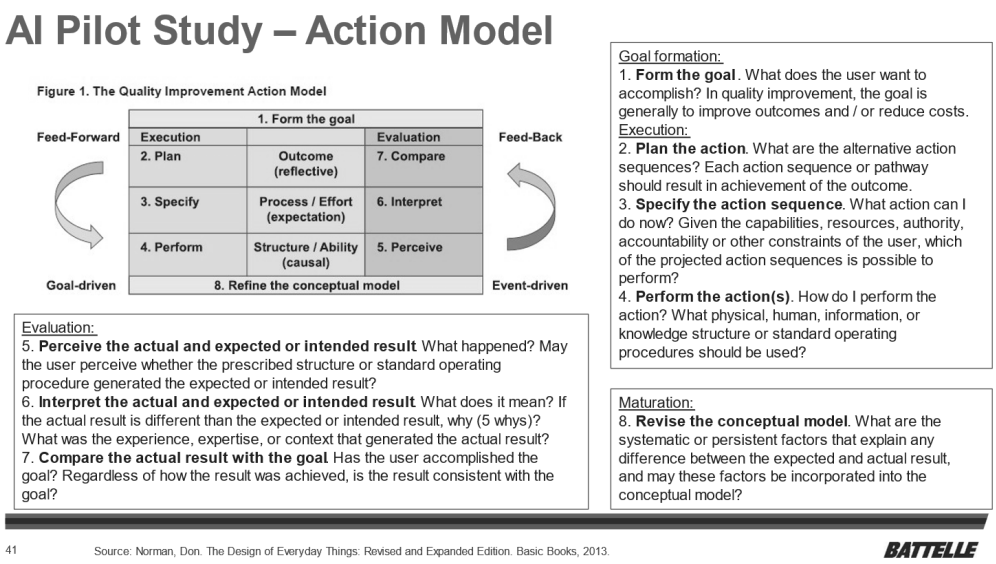
We have a model to accomplish that, or I should call it a quality improvement action model. We want to understand. That's sort of our

controller is that model. We want to understand what kind of actions make that controller unsafe. So what's the context that makes it unsafe? What are the triggers necessary in order for that model to work? And then, identify scenarios in which our quality improvement model doesn't work.

So ultimately, you know, kind of the big picture here is that when we have a measure, we want to know very explicitly what it is that folks need to do in order to perform well on that measure. "What it is that is preventing them from performing well on that measure, what they need to do in order to perform better on the measure, and what are the barriers that are preventing them from doing so?"

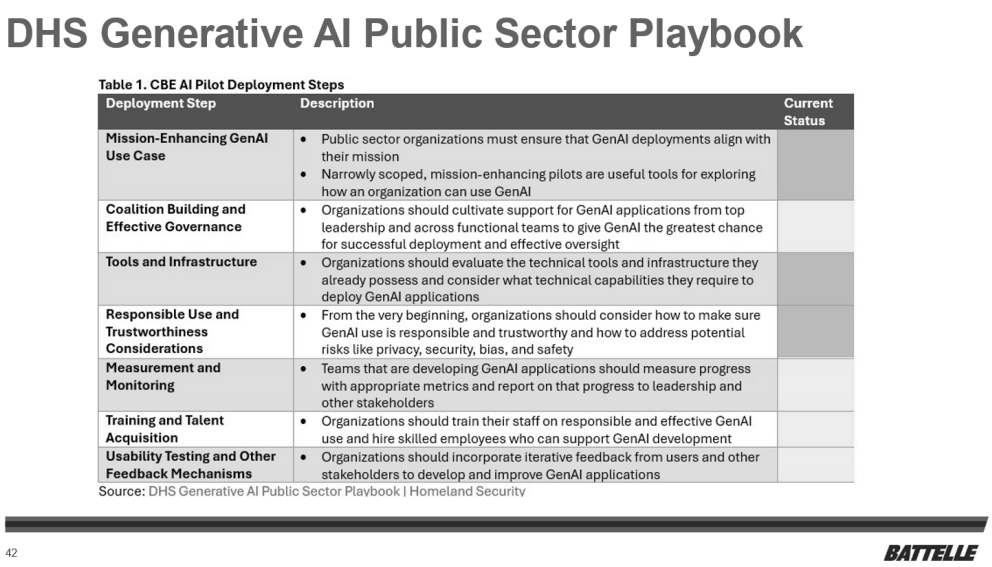
So if we can address all four of those questions, then we feel like our measurement system is operating as intended. So I'll just wrap up there.

[SLIDE 41]



GEPPERT: So the next slide is a description about what the components of the action model are, how it works, and what it is that you need to be able to articulate. AI can help us to sort of articulate these different aspects to the model.

[SLIDE 42]



GEPPERT: And then this is a playbook that was developed by the Department of Homeland Security (DHS) about how one implements these sort of generative-AI type solutions. It needs to be mission-focused. There needs to be a governance organization. You need to have the tools and infrastructure in place. You need to make sure that you're ensuring trustworthiness. You need to measure and monitor. You need the talent in order to be able to operate it. And then you need to conduct usability testing. So all of these things are sort of informing our thoughts about how to do this going forward.

[SLIDE 43]

DHS Generative AI Public Sector Playbook

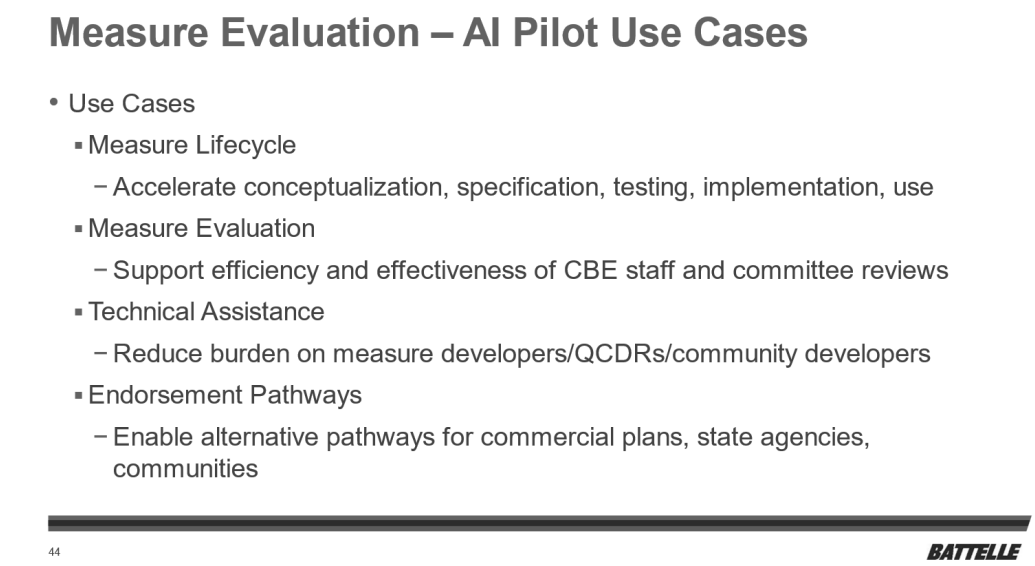
Table 1. Key Performance Indicators for AI Pilot

Domain	Metric	
Percentage of claims by source	Number	Percentage
• Provided claims only		
• CMO claims only		
• Both Provided and CMO claims		
• **Total**	Number	Percentage
Status of CMO claims (SME review)		
• Arguable true		
• Speculative or Arguably false		
• **Total**	Number	Percentage
Claim Status Justifications (SME review)		
• Agree		
• Neutral		
• Disagree		
• **Total**		

43 BATTELLE

GEPPERT: And then this is sort of how we thought about developing a set of key performance indicators (KPIs) that we could use to track this over time.

[SLIDE 44]



Q&A
[SLIDE 45]

GEPPERT: So I do want to get to some of the questions. So we've looked at a couple of examples. In terms of informing data abstraction, I think it does in the sense that—essentially though the question that came up just a few days ago about being better able to understand what are the steps involved with abstracting data. There are some modeling approaches that people are starting to use to be more explicit about the

steps that are needed, the resources that are needed, and the processes that are needed to abstract the data. So that is very much part of our burden assessment and how we think about trying to be more explicit about both burden and benefit.

There's a question about whether we've applied the methodology to the patient-reported outcome measure. We haven't, and we would love to do that. If it's something of interest to you, please reach out to us.

There's a specific sort of request to apply it to a specific measure. We'd love to do that. If you have measures that you're interested in and would like us to take a look at, we'd be more than happy to do that.

In terms of the literature review, I would recommend or suggest that if you're able to attend the webinar that we're going to do in a couple weeks where we're specifically talking about literature reviews and how we can use these AI tools to perform literature reviews — that's what we're going to be focused on at our next webinar event, the next webinar. I think it's on the 17th.

So then, you know, how this could be used. I think this really, I think ideally, we want to be able to use these tools as early as possible, right? Like as early in the measure development lifecycle as we can to be able to sort of describe the measure, understand the measure, you know, and map out our plan for testing the measure. I think that's really the intent. It's to try to move these tools as early as possible in the measure lifecycle, and then hopefully to accelerate that lifecycle as much as we can, and to reduce as much burden as we can.

So that's really sort of the *goal*. It's to help guide measure development. So then by the time it gets to a measure evaluation sort of process, you know, a lot of these things have already been brought forward and discussed and resolved, which kind of speaks to how it might influence the measure development process. We rely, for example, a lot on the technical expert panels (TEPs) as part of our measure development process. I think they could play a critical role here, right?

As the SME reviewers of some of this output and how the TEPs review some of this AI output and evaluate it, and to identify the claims that are going arguably true or speculative — that could be a great role for the TEPs going forward.

The question about LLMs and clinical data and claims data. There is sort of a growing literature about that, you know, using LLMs to extract data from unstructured fields. I think that is certainly happening, and it's evolving. I think that the most recent things that I've seen still involve some human review about—again, it's sort of a one left, one step higher where it's not so much replicating what LLM did. But then one of the nice things about LLMs is that they provide reasons and rationales. You can evaluate those reasons and rationales for plausibility, and I still think there's a step of evaluating the reasons and rationales for plausibility.

So then there's still some human input in terms of the data extraction. So you've asked some great questions. I'll stop there, but please reach out to us and we'd love to talk to you more about this.

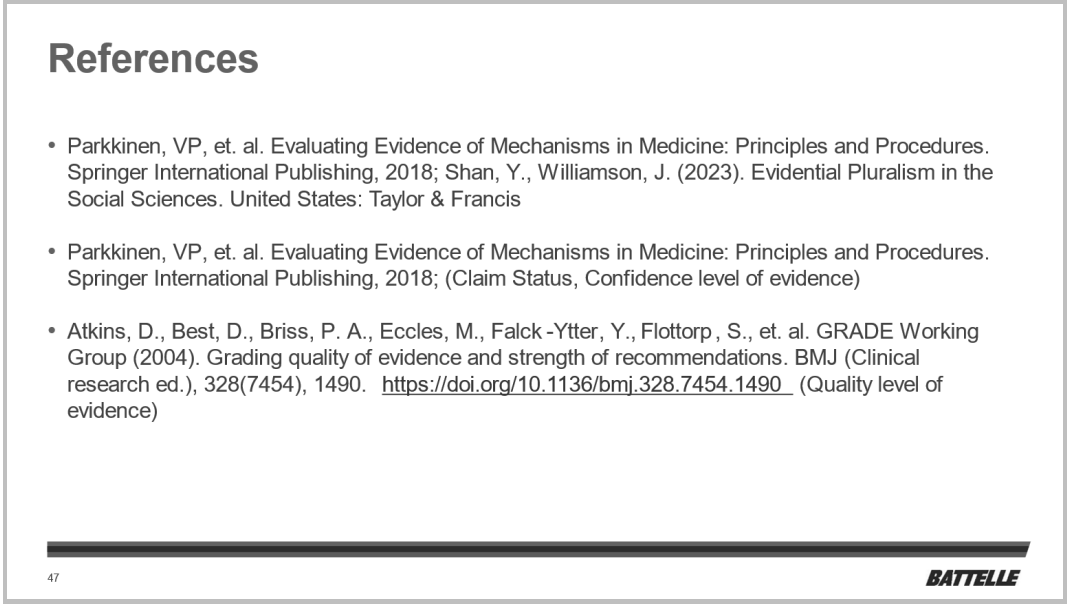
[SLIDE 46]

References

- https://nam.edu/programs/value -science-driven-health-care/assessing-meaningful-community-engagement (CBE Strategy)
- Pawson, R., Tilley, N. (1997). Realistic evaluation. United Kingdom: SAGE Publications (context - mechanism-outcome ontology)
- Thissen D, Wainer H. Test scoring. Lawrence Erlbaum Associates, 2001; Messick, S. (1989b). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13 -103). New "York: Macmillan; Standards for Educational and Psychological Testing. (2014). United States: American Educational Research Association (modern validity theory)
- National Institute of Standards and Technology (2022) Engineering Trustworthy Secure Systems (NIST SP 800-160 Vol. 1 Rev. 1) https://csrc.nist.gov/pubs/sp/800/160/v1/r1/final ; International Organization for Standardization. (2019). Systems and software engineering — Systems and software assurance – Part 1: Concepts and vocabulary (ISO Standard No. 15026 -1:2019). https://www.iso.org/standard/73567.html (assurance cases)

46 BATTELLE

[SLIDE 47]



[SLIDE 48]

MODERATOR: Thank you so much, Jeff, and thank you everyone for joining today's Information Session. The slides will be available on the

MMS Hub following the session, and we will have the video and Q&As available in the coming weeks. Thank you all so much.

WEBINAR CONCLUDES