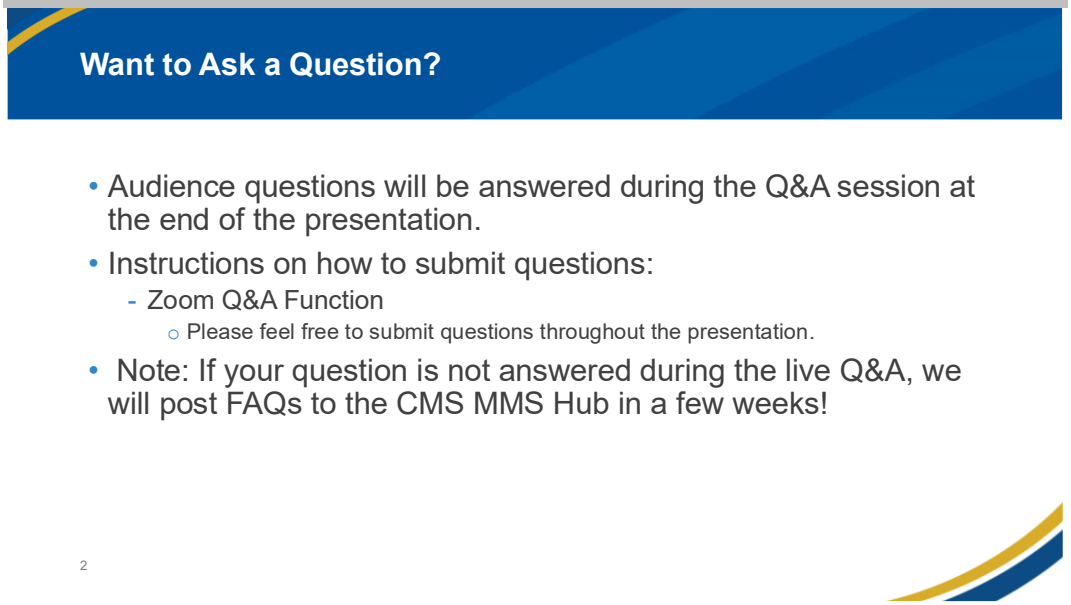
2026 MMS Information Session:
Strategies & Efforts
For Improving Diagnostic Performance Transcript

[SLIDE 1]

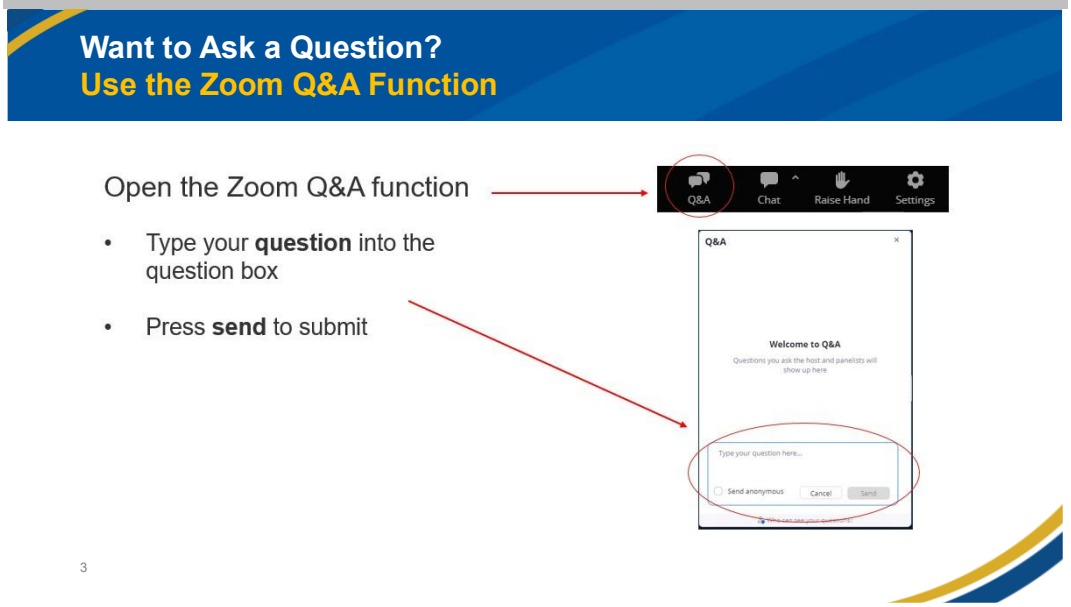
GHUNNEY: Good afternoon. Thank you for joining us for today's CMS Measures Management System (MMS) Information Session: *Strategies and Efforts for Improving Diagnostic Performance*. My name is Aya Ghunney, and I work for Battelle in support of MMS. This webinar is just one aspect of MMS education and outreach, and I invite you to visit the MMS Hub website to learn more about MMS and how to get involved in quality measurement.

[SLIDE 2]



GHUNNEY: So throughout today's presentation I encourage you to submit questions using our Q&A feature which you're going to find near the bottom of your screen. We'll be addressing questions during the live Q&A at the end of the presentation. And then following the meeting we'll share the meeting materials, including the questions and answers (Q&As) that we went through today on the MMS Hub website.

[SLIDE 3]



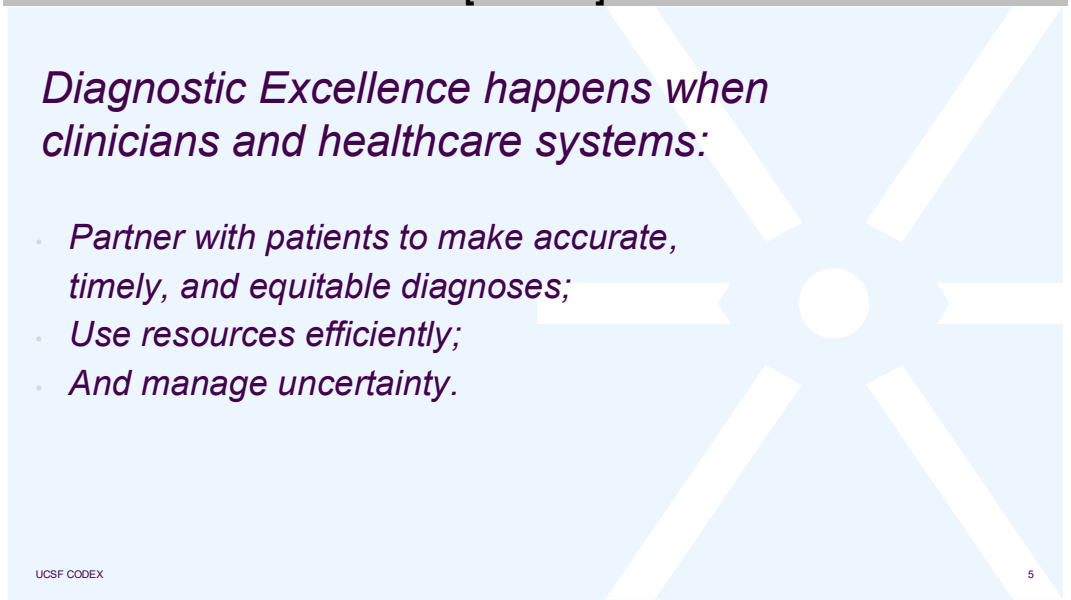
GHUNNEY: Okay, so this is just showing you exactly where that Zoom Q&A function is. Please feel free to send in your questions as we start the discussion.

[SLIDE 4]

GHUNNEY: So today we're very excited to welcome Dr. Sumant Ranji and Dr. Ben Rosner from the University of California-San Francisco Coordinating Center for Diagnostic Excellence, or CODEX. So Dr. Ranji is a national expert in patient safety with a particular focus on diagnostic excellence (DxEx). He's a professor of clinical medicine and the director of CODEX. We're also joined by Dr. Ben Rosner who's a professor, faculty lead for the CODEX Action Incubator, a hospitalist, informaticist, and diagnostic performance feedback researcher. And with that, I will hand it over to Dr. Ranji to get us started.

RANJI: Thank you so much, Aya, and I want to thank Battelle and CMS for the opportunity to present to you today about our ongoing work on developing measures of diagnostic excellence (DxEx) in the era of artificial intelligence (AI). So Ben and I will be discussing our ongoing work, which is still in progress, and we're very excited to present this work to you today with the goal of helping everyone understand how we as a center are thinking about measuring diagnostic excellence (DxEx), and we would also love your input on our work which is in progress right now.

[SLIDE 5]



RANJI: So I'm going to start by giving a little bit of an overview of diagnostic excellence (DxEx) and discuss our role at CODEX, which is our UCSF Center for Diagnostic Excellence, and the role that we play in diagnostic excellence (DxEx) and trying to improve diagnosis overall. And then I will transition into discussing the work that Ben is leading with our Action Incubator, which is really centered around trying to improve the measurement of diagnostic excellence (DxEx) with a particular focus on diagnostic excellence in the era of artificial intelligence (AI).

So when we talk about diagnostic excellence (DxEx), what we are referring to is the ability of a health system and the clinicians who work within it to provide patients accurate, timely, and equitable diagnoses while also using resources efficiently and managing uncertainty.

At CODEX we very much view the process of diagnosis as a partnership with patients, and as we'll discuss, part of our work centers around understanding how patients themselves can evaluate the quality of

diagnosis and participate in the process of diagnosis. I would also emphasize here that diagnostic excellence (DxEx) is about more than just preventing diagnostic error. From a patient safety standpoint diagnostic error, as I'll discuss, is a well-recognized patient safety issue that has evolved in different ways over the last several years. But when we talk about diagnostic excellence (DxEx), of course, preventing errors is an important part of that, but we really view diagnostic excellence as more than just the prevention of error. It involves the achievement of a system in which all patients are able to receive accurate, timely, and equitable diagnoses.

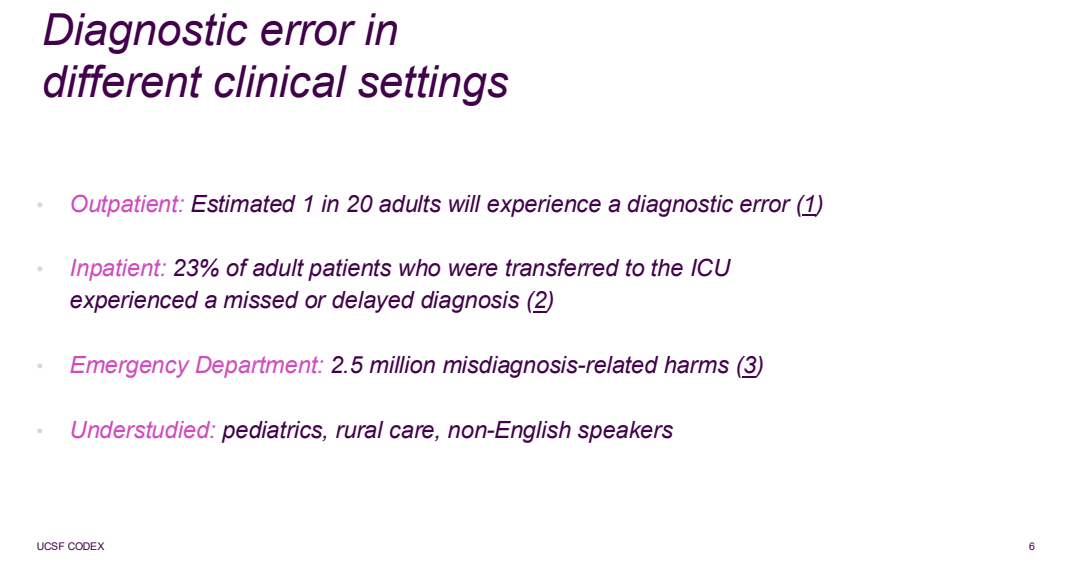
The domains of diagnostic excellence (DxEx) involve accuracy of diagnosis. So was the diagnosis correct or not, or as correct as possible? But it also includes timeliness. So did the patient receive a diagnosis in a timely fashion in order to be able to benefit from appropriate treatment?

And another domain that's very important is communication. Even if the clinician and the health system do everything correctly in terms of establishing a diagnosis, if that diagnosis is not communicated to the patient in a way that they understand — in a way that helps the patient understand both the implications of the diagnosis and the resulting treatment plan — then the system has failed the patients, as well as these domains.

The use of appropriate resource utilization is very important both from a patient standpoint, because inefficient use of resources can lead to financial toxicity for patients. And as well, from the health system standpoint, even if a diagnosis is reached by an inefficient process, that incurs great costs to the health system overall.

Finally, I'd be remiss if I didn't discuss aspects of equity as a component of diagnostic excellence (DxEx). There are well-recognized disparities in diagnosis as there are in other patient safety issues. So I want to spend a little bit of time talking about CODEX and our role within the diagnostic excellence ecosystem.

[SLIDE 6]



RANJI: So before talking about CODEX, I wanted to spend a minute just talking about the epidemiology of diagnostic error and what we know about the risk for diagnostic error in different clinical settings. So a great deal of foundational work has been done really within, I would say within the last 5-10 years to establish the prevalence of diagnostic error and the severity of harm in different clinical settings. So some of the key studies are listed here, but overall it's estimated that in their lifetime one out of every twenty (20) adults will experience a diagnostic error, either in the inpatient setting or in the outpatient setting.

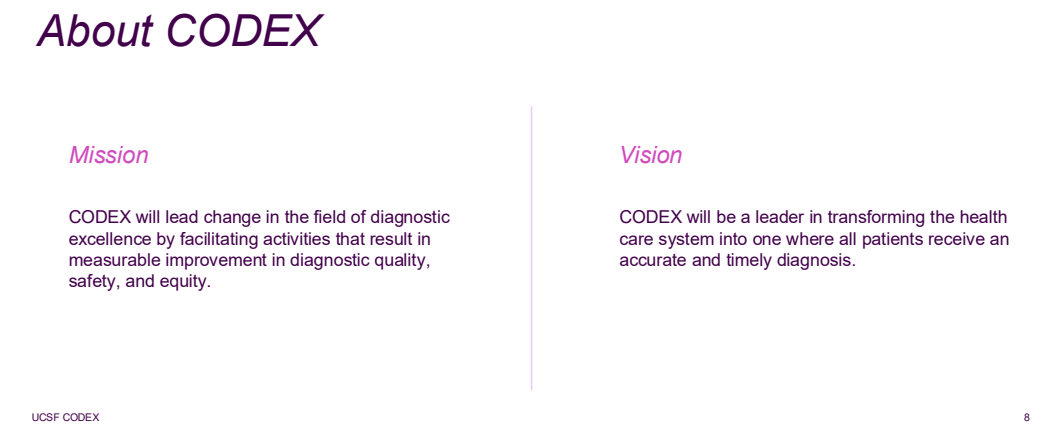
Some recent work done by some of my colleagues at UCSF established that in hospitalized medical patients 23% of adult patients who experienced a clinical deterioration while hospitalized — that is patients who died or were transferred to the ICU — actually experienced a missed or delayed diagnosis. And then in patients who had a mixed or delayed diagnosis while hospitalized, they had significant clinical consequences for them in terms of increasing their risk of severe morbidity or mortality.

Diagnostic errors have also been characterized in the emergency department (ED), but there are a few areas in which we don't know as much as we should. Most studies of diagnostic errors have taken place in health systems that are predominantly in urban or suburban areas. So we really don't know that much about diagnostic error, for example, in rural settings. There's a lack of research on the intersection between social determinants of health (SDOH) in diagnostic error, particularly around patients whose primary language is not English. And while there are some good studies around diagnostic error in pediatrics, I would say there's probably less data there than there is in adult patient populations.

[SLIDE 7]

RANJI: So I'm going to transition now to talking about CODEX and what our role is within efforts to improve diagnosis.

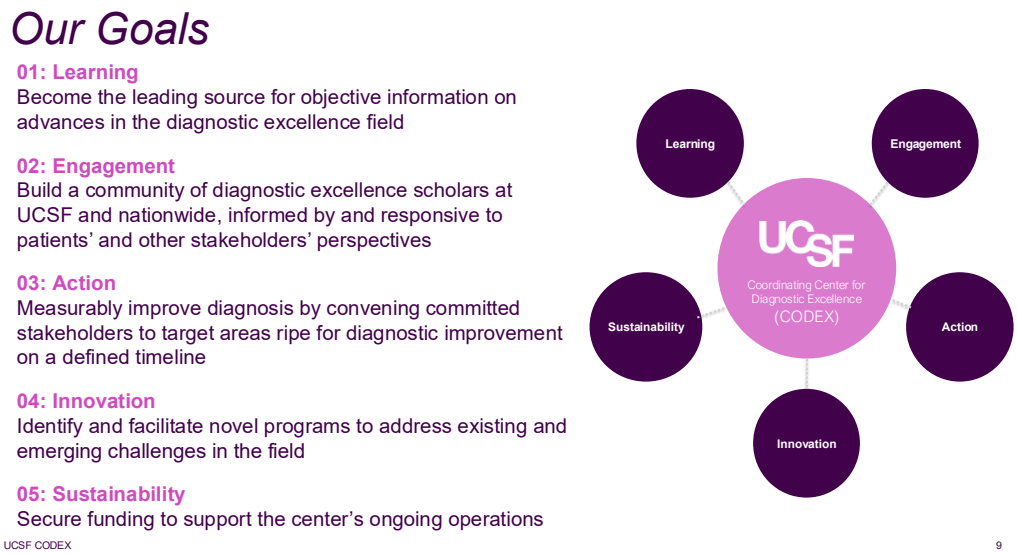
[SLIDE 8]



RANJI: So CODEX was established in 2024, thanks to a generous grant from the Gordon and Betty Moore Foundation. Our mission is to lead change in the field of diagnostic excellence (DxEx) by facilitating activities that result in measurable improvement in diagnostic quality, safety, and equity. Our vision is to transform the healthcare system into one where all patients receive an accurate and timely diagnosis.

Our work in CODEX is divided into several domains which I'll talk about. We view our role within CODEX as partly research, but also supporting the overall field of diagnostic excellence (DxEx), which definitely includes improving the measurement of diagnostic excellence.

[SLIDE 9]

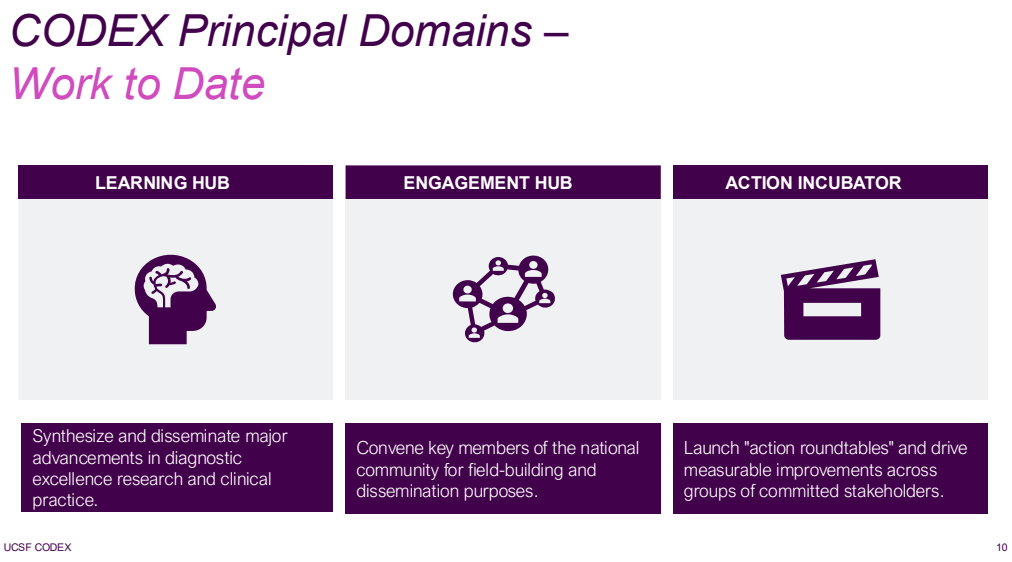


RANJI: So our main goals are listed here. So I'll talk about each of these in a little bit more detail, but a big part of our work involves learning. So disseminating the latest research and innovations in the diagnostic

excellence (DxEx) field. In this area because of the rapid promulgation of artificial intelligence (AI) within healthcare — which is increasingly being used as a diagnostic aide — we are centering a lot of our work around assessing the literature both published and unpublished around the use of artificial intelligence in improving diagnosis and trying to make sense of all the studies which are coming out literally week by week on the use of artificial intelligence in diagnosis in particular.

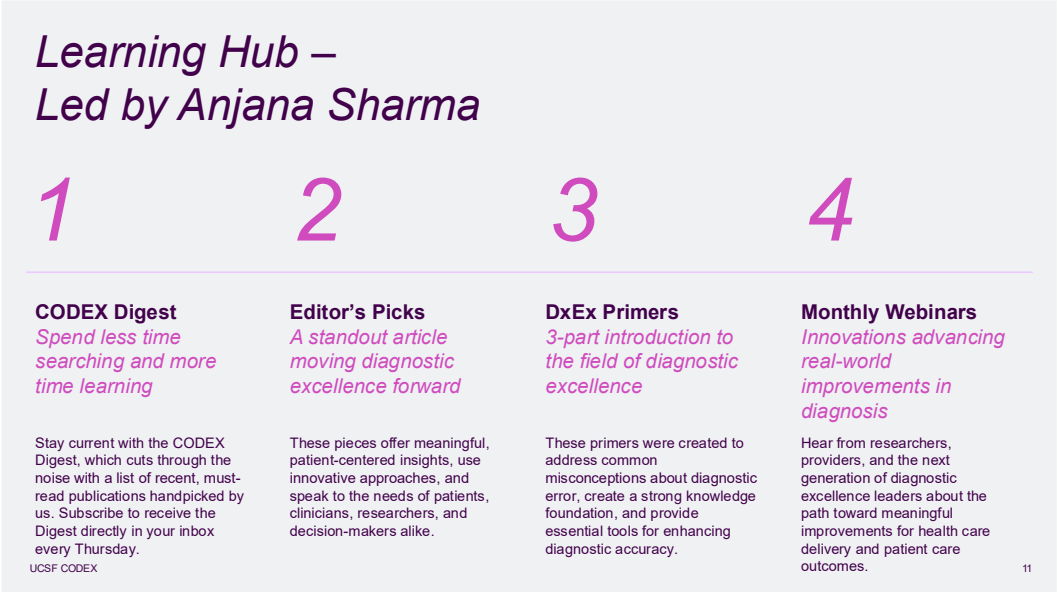
As I said, a big part of our responsibility is to support the diagnostic excellence (DxEx) field as a whole, which involves both capacity building and expanding the field beyond its traditional areas. And then what will be the primary focus of our talk today is on our Action Incubator work, and the Action Incubator which is the work the band has been leading, has the goal of measurably improving diagnosis by bringing together stakeholders to target areas ready for diagnostic improvement on a defined timeline.

[SLIDE 10]



RANJI: Our three main domains are the learning hub, the engagement hub, and the Action Incubator. So I'm just going to talk a little bit about each of those before we transition into the measurement-oriented part of the talk.

[SLIDE 11]



RANJI: The learning hub work is led by our colleague, Anjana Sharma, who's a family physician who works with me at San Francisco General Hospital. Our goal here is really to be a source for unbiased information for learning about advances in diagnostic excellence (DxEx). We put out a weekly digest which summarizes 10-12 key studies in the area of diagnosis, particularly focusing on artificial intelligence (AI). This has gotten really good response from our audience, and I think is a great way to stay on top of the latest developments in diagnostic excellence (DxEx).

We also run a monthly webinar series entitled "Advances at the Intersection of Diagnosis and Artificial intelligence (AI)," where we try to

bring together the informatics community and the patient safety community to discuss the latest applications of AI to improving diagnosis, and bring in discussions who are thinking about AI implementation from a variety of perspectives.

[SLIDE 12]

RANJI: Here's an example of our digest which comes out weekly. This QR code will take you to our CODEX website where you can sign up to receive the digest. What we do is we identify 10-12 articles per week. They're curated by our expert cybrarian who then also writes a brief summary of each article, allowing for understanding of where the articles' findings fit into the overall literature around diagnostic excellence (DxEx).

[SLIDE 13]

RANJI: In addition to our learning work, we are also doing as much as we can to support the growing diagnostic excellence (DxEx) field. We run an annual conference, the diagnostic excellence conference, which most recently took place last fall in Ann Arbor, Michigan. This is an annual thing. We'll be announcing the site for the diagnostic excellence 2026 meeting and the dates very shortly, probably within the next week or so.

We also are partnering with the National Academy of Medicine (NAM) to support their forum on advancing diagnostic excellence (DxEx), as well as their fellowship in diagnostic excellence. We have an ongoing partnership with patients for patient safety, which is a leading patient-led organization in the patient safety field who have a strong interest in improving diagnosis and are leading a lot of work around patient engagement, patient advocacy, and measuring patient perspectives on improving diagnosis.

[SLIDE 14]

RANJI: So I'm going to turn now to talking about the current state of measurement of diagnostic excellence (DxEx). So diagnostic excellence and diagnostic error is an area that within patient safety traditionally has had some challenges in terms of measurement. Among these challenges include the lack of diagnostic gold standards. So even for diseases that are quite common, many of them often don't have a single gold standard either an imaging test or blood test which allows us to definitively say, "Does the patient have this diagnosis or not?" So many things that, you know, either Ben and I see in the hospital or see in primary care, the final diagnosis is really a combination of objective testing and clinician judgment. That alone makes it labor-intensive as well as somewhat challenging to establish just the denominator of patients who have a certain diagnosis.

Another challenge which is an ongoing challenge is the lack of accurate electronic health record (EHR) data around key aspects of diagnosis. Just

as an example, symptom data is not reliably captured within most electronic health records (EHRs) currently. So from a research standpoint and from a quality improvement standpoint, if we don't know exactly what symptoms a patient experienced or when they started experiencing those symptoms, it makes it very difficult for us to be able to judge "was the diagnosis made in a timely fashion," even if we can say, "was the final diagnosis accurate?" If we don't know when the patient started experiencing symptoms, or how characteristic their symptoms were, it's hard for us to say if the diagnosis was made in a timely fashion and/or in the most efficient fashion.

EHR data reliably, or fairly reliably includes the patient's actual diagnoses, but it often takes a fair amount of retrospective work to determine when was the diagnosis actually established and how did the differential diagnosis evolve over time.

Another challenge in measuring diagnostic excellence (DxEx) has been the lack of standardized measures of the patient experience as it pertains to diagnosis. So while there's been great progress in measuring patient experience of other patient safety issues, as well as patient engagement overall in the process of care, we're still working on developing both patient-reported outcome measures (PROMs) as well as patient-reported experience measures (PREMs) for diagnostic excellence. As I'll discuss, there has been a good amount of progress in that area, but really just within the last few years.

So those are challenges I think that have limited our ability to both measure diagnostic excellence (DxEx) both on a disease-specific basis, as well as from an overall standpoint. But, as I'll discuss, there has been considerable progress made within the last couple of years. So just

highlighting on this slide a few of the advances that have been made, and of some of the measures that are being implemented both from the federal and nonfederal standpoint.

So I'm sure there are many people on this call who know more about this than I do, but CMS has implemented a number of both measures and associated with improvement activities to support improving diagnosis over the last few years. Some highlights of these, as many of you no doubt are aware, within the MIPS program there are improvement activities that center around improving diagnosis that health systems can report. These include improving the diagnosis of cognitive impairment, closing the referral loop between specialists and primary care clinicians, as well as newer programs around reporting measures that organization are taking to improve the safety of artificial intelligence (AI), which often is being used for diagnostic purposes, and of course, the 2025 MUC List includes two diagnostic excellence (DxEx) measures, which I'll discuss in detail a little bit later.

From the standpoint of the Agency for Healthcare Research and Quality (AHRQ), as the AHRQ is undertaking revisions to their Quality Indicators Program, there are actually 14 different measures of diagnostic excellence (DxEx) that are under consideration for incorporation to the next iteration of the AHRQ QIP program. These are almost all disease-specific measures, but in large part I think are derived on advances both in electronic health records (EHRs) as well as the traditional measure development process.

Finally, from a nonfederal standpoint, the LeapFrog Group has added questions to its annual rating surveys to assess what actions hospitals are taking to reduce harm to patients from diagnostic errors. These are

focused more on structural measures rather than process or outcome measures, but the fact that LeapFrog is starting to include these I think highlights the salience of diagnostic errors as a patient safety problem.

[SLIDE 15]

Current State of Measurement of Diagnostic Excellence

Gordon and Betty Moore Foundation Measurement Portfolio

Pre-Rulemaking and Use in Federal Quality Programs and Consensus-Based Entity Endorsement:

Measure Title	Measure Developer	Current Status
CVD Risk Assessment Measure	University of California, Irvine	Implemented in MIPS
Appropriate Germline Testing for Ovarian Cancer Patients	American Society of Clinical Oncology	Implemented in MIPS
Delay of Venous Thromboembolism in Primary Care	Brigham and Women's Hospital	Included in the proposed rule for use in the MIPS program
Follow-Up on Abnormal Screening Mammograms for Breast Cancer	Brigham and Women's Hospital	Included on 2025 MUC List
Follow-Up on Positive Stool-based Tests for Colorectal Cancer	Brigham and Women's Hospital	Included on 2025 MUC List

CBE ID	Measure Title	Measure Developer	Endorsement Status
4440e	% hospitalized pneumonia patients with chest imaging confirmation	University of Utah	Endorsed
3671/3690	Inappropriate dx of CAP Inappropriate dx of UTI	University of Michigan	Endorsed
3617	Provider Level Continuity of Care	American Board of Family Medicine	Endorsed
3749e	Delay of Venous Thromboembolism in Primary Care	Brigham and Women's Hospital	Endorsed

RANJI: Many of the advances that have taken place in measuring diagnostic excellence (DxEx) are thanks to investments made by the Gordon and Betty Moore Foundation who founded CODEX and have invested quite a bit in their diagnostic excellence (DxEx) program over the last several years. So the Gordon and Betty Moore Foundation had a longstanding diagnostic excellence portfolio, and a good portion of that was focused on funding organizations to develop measures of diagnostic excellence.

I'm not going to summarize everything that was done under the Moore Foundation's supervision, but these are some of the highlights of the work

that was carried out by organizations that were funded by Moore, that have made it through the approval process for quality measures.

So, as you can see here, there are some measures that have actually been implemented in the MIPS program or included on the MUC List. There are multiple other measures that have made it through the consensus-based endorsement process but have not yet been implemented, and so these are programs which, as I said, were funded by the Moore Foundation and over the last several years have been validated and either implemented or endorsed.

[SLIDE 16]

Current State of Measurement of Diagnostic Excellence

Gordon and Betty Moore Foundation Measurement Portfolio

American College of Emergency Physicians CEDR Registry Use (2026 Implementation):

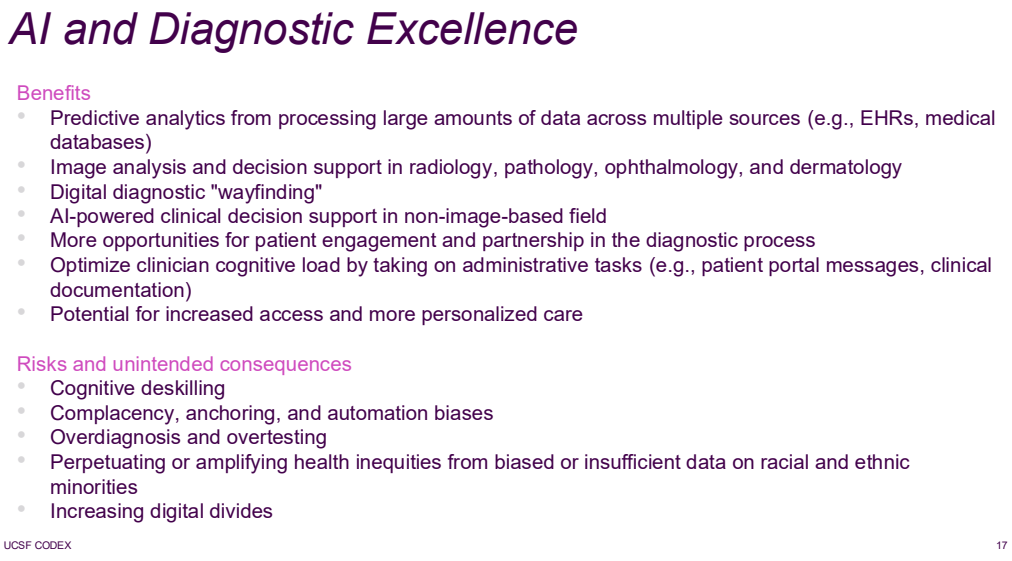
Measure Title	Measure Developer
Diagnostic Utilization – Proportion of all ED visits with chest CT ordered to investigate PE	Yale University, ACEP
Diagnostic Yield – Proportion of chest CTs ordered with Acute PE	Yale University, ACEP
Diagnostic Opportunity – Proportion of ED diagnostic opportunities 0-7 days prior to ED PE with PE diagnosis	Yale University, ACEP

UCSF CODEX 16

RANJI: In addition to these, the American College of Emergency Physicians (ACEP) is in the process of implementing diagnostic utilization measures focused around acute pulmonary embolism (PE) and these are metrics which organizations can choose to report as part of the MIPS program. So while, as I said, measurement of diagnostic excellence

(DxEx) is somewhat challenging, there has been a considerable amount of work done over the last several years, in large part thanks to the Moore Foundation, and that has resulted in the development of several measures that are now being widely used to measure various aspects of diagnostic excellence.

[SLIDE 17]



RANJI: But that brings us to the current state, which is what has really changed over the last couple of years is the more widespread implementation of artificial intelligence (AI) in healthcare and its implications for diagnostic excellence (DxEx). So AI in various forms, whether it involves machine learning or whether it involves deep neural networks (DNNs) — or increasingly the use of large language models (LLMs) — is actually quite widely implemented throughout healthcare.

I think it's safe to say that the areas where AI broadly speaking is used most commonly are in image-based fields such as radiology and

pathology. In radiology in particular the majority of hospitals in the country are now using at least one AI-based application, and the FDA has approved over 800 different AI-based diagnostic decision support tools focused on improving diagnosis in radiology.

So if a patient comes into the hospital and it's suspected that they may be experiencing an acute stroke, more likely than not when they get a CT scan or when they get an MRI, that scan is being read both by a radiologist as well as by an artificial intelligence (AI) application. The radiologist has the ability to access the AI read of the scan in order to arrive at a final diagnosis of whether or not a patient has had a stroke.

There are similar examples in other image-based fields, as I said, including pathology and dermatology. It's quite common in pathology for slides to receive an AI read in conjunction with the pathologist read, which allows for mutual crosschecking, and ideally a more accurate identification of disease processes.

So AI has quite a bit of potential benefit for improving diagnosis. The fact that AI is able to access large amounts of data and be trained on gigantic databases across multiple sources, it definitely allows for the potential to develop both predictive analytics and models to more accurately diagnose common conditions.

As AI has progressed there is increasing interest in the use of large language models (LLMs) to help diagnosis in more common clinical settings. In multiple simulated studies large language models (LLMs) have actually been shown to have diagnostic capabilities similar to, and in some cases superior to clinicians. Now I would caution that these studies have not yet been replicated in actual clinical practice, but there is a lot of

ongoing research in that area. And in addition to the research that's going on, many of the leading AI companies as well as electronic health record (EHR) developers are in the process of developing decision support systems (DSSs) that are based on AI to help clinicians improve their diagnostic capabilities.

Probably the leading example of day-to-day implementation of AI outside of image-based fields are artificial intelligence-based scribes, and some of you may have experienced this just in the course of being a patient or going to a clinician. In many settings, but mainly in the outpatient setting, the implementation of AI scribes is pretty widespread. This is a technology that really helps clinicians from the standpoint of being able to spend less time just entering data into the computer. It allows for more natural interaction with the patient, and we hope will help optimize a clinician's cognitive load and allow them to focus on the patient rather than on data entry, and hopefully free up time for the more cognitively intensive aspects of being a clinician, including the diagnostic process.

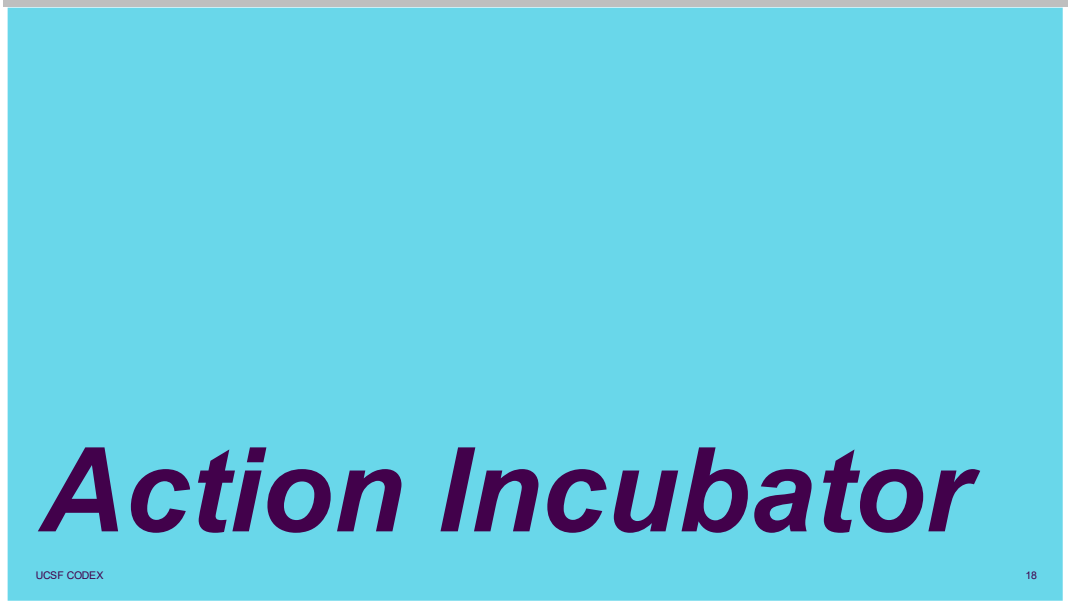
So there's a lot of excitement about AI. It's being implemented fairly widely in certain fields, and there's rapid expansion in other clinical areas as well. A lot of these areas hopefully it's clear how this impacts diagnosis in image-based fields, but our interest in CODEX is in studying how the implementation of AI broadly affects the diagnostic process, both in terms of benefits and potential unintended consequences. So the unintended consequences of artificial intelligence (AI) implementation for diagnosis are listed here.

I would highlight "cognitive deskilling" which refers to when if clinicians become overly reliant on AI or are trained in an AI world — they may become overly reliant on AI to help them either with diagnosis or treatment

plans — which may result in their inability to actually take care of patients if they don't have access to AI. So the issue of cognitive deskilling is a very real one that is a huge topic in the medical education world. As AI is implemented more broadly, how do we make sure that clinicians — be they physicians, nurses, advanced practice providers — are able to learn core clinical skills without having to rely on AI at all times? There are other potential unintended consequences as well, including overdiagnosis and overtesting based on how large language models (LLMs) in particular work, which have a great ability to synthesize data but also may bias towards rarer or more uncommon diagnoses.

So as we in CODEX take a step back and think about the diagnostic excellence (DxEx) world as a whole, it's clear that AI is having a big impact and probably will continue to have a big impact for the next several years. So as a result of that, we have spent a lot of time thinking about how we can try and ensure that as AI continues to advance that it actually achieves the potential that it is.

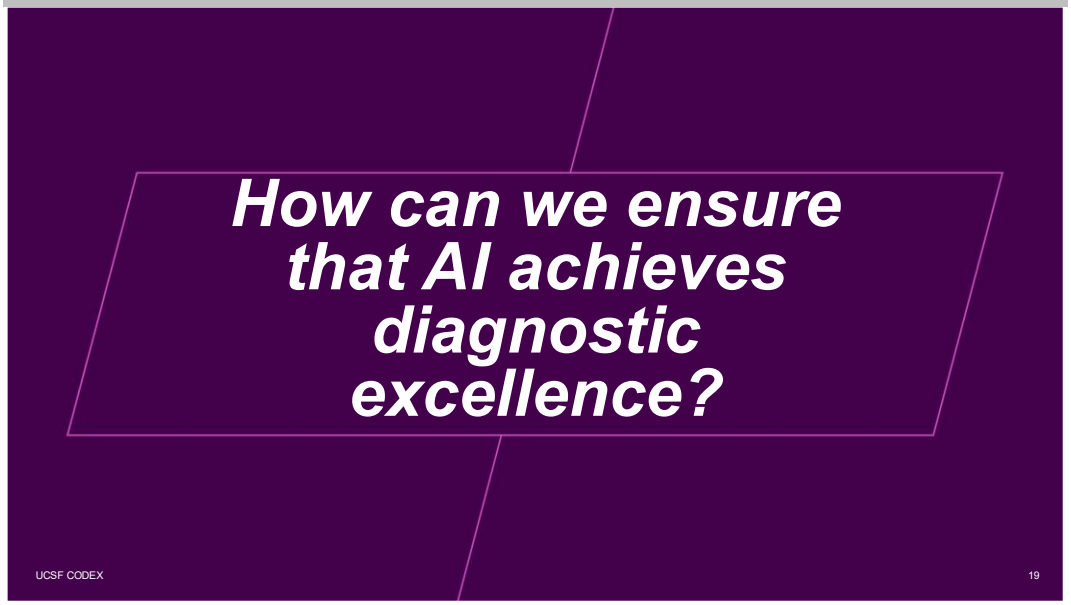
[SLIDE 18]



RANJI: And that is where our Action Incubator work comes in. So I will turn things over to my colleague, Ben Rosner, who has really led our Action Incubator work. So just to give you a highlight of what the Action Incubator is, this is really as the name says, our "action-oriented arm" of CODEX. It's through the Action Incubator that we hope to achieve measurable improvement in diagnosis by focusing on key areas in diagnosis that either have good measurement strategies or need good measurement strategies and are ripe for improvements on a defined timeline.

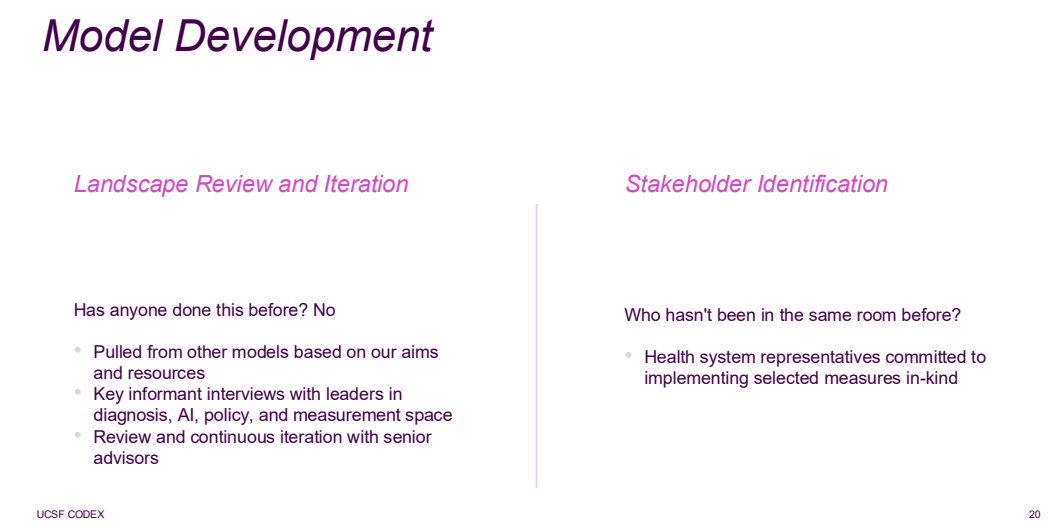
So we conducted our Action Incubator quite recently and are in the process of our collaborative work that is following the actual in-person meeting. So let me turn things over to Ben who will take us through our ongoing work. And then after that we'll definitely have time for questions and reactions.

[SLIDE 19]



ROSNER: Thank you. As Sumant had mentioned earlier, the goal of the Action Incubator is really to try to amplify research by translating it into tangible next steps. And the idea with CODEX for an Action Incubator is not a one-off. In fact, we are interested in having recurring Action Incubators where we bring together people in the area to try to then move things into actionable next steps. And, as Sumant mentioned, we held our first one this past September. I'll share with you a little bit about how we developed the model for what it is, what our thought process is, and also at least where this first Action Incubator is heading before we think about where the future ones will move as well.

[SLIDE 20]



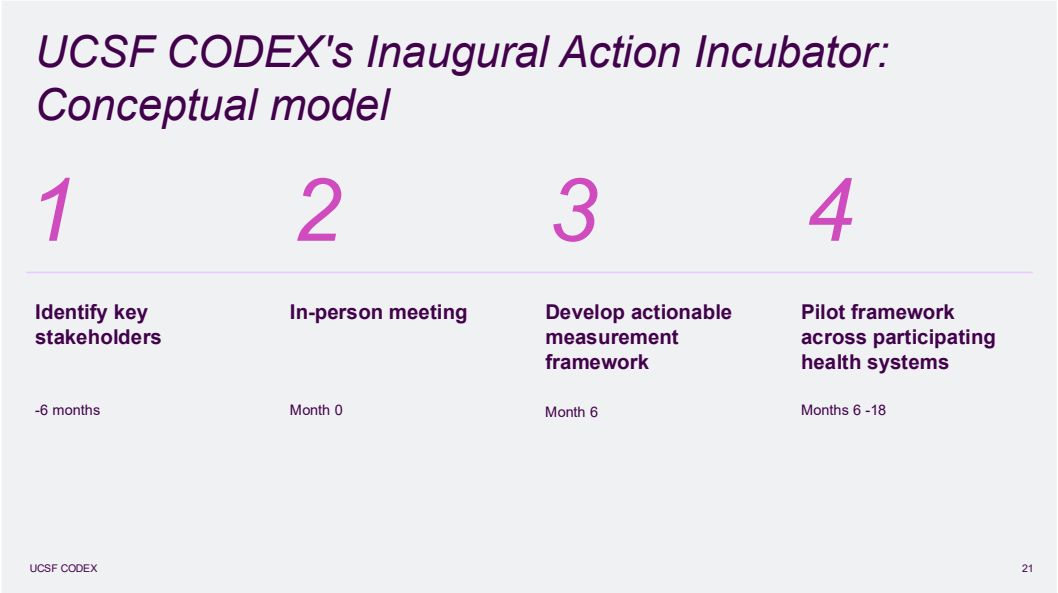
ROSNER: As we thought about some of the models for what is Action Incubator, and what is in essence the optimal way to carry one of these out, we spent time doing a landscape review to try to examine are there other examples of successful Action Incubators that are trying to translate in essence research and concept into action. There are a few examples that have been done before, but none specifically to diagnosis. However, some of these models did give us ideas for approaches for best practice.

So, for example, we interviewed approximately 20 thought leaders from around the country and asked them questions such as, "What is in your opinion a good or optimal product of an Action Incubator? Is this a white paper, or is this a research article, or is this putting things into implementation in health systems, for example?" We also asked them, "What is the optimal size of an Action Incubator? How many people should be involved?" You can imagine there might be a sweet spot where

you have enough stakeholders involved where you have a variety and breadth of concepts and ideas, but you don't dilute that experience by too many people where your engagement drops off.

So after interviews with about 20 national thought leaders, we started to center around the concept of an in-person Action Incubator — one to two days — and limited to approximately 30-35 participants so that there would be high engagement and participation. And in so doing, we also recognized that there was actually quite a bit of a need for some diversity in approach. You don't want to just bring those to the table who are in let's say only research, or those to the table who are only patient safety, and also those to the table who are only from one or two industries. We actually came to learn a great diversity of stakeholders was actually necessary to move something like this forward.

[SLIDE 21]

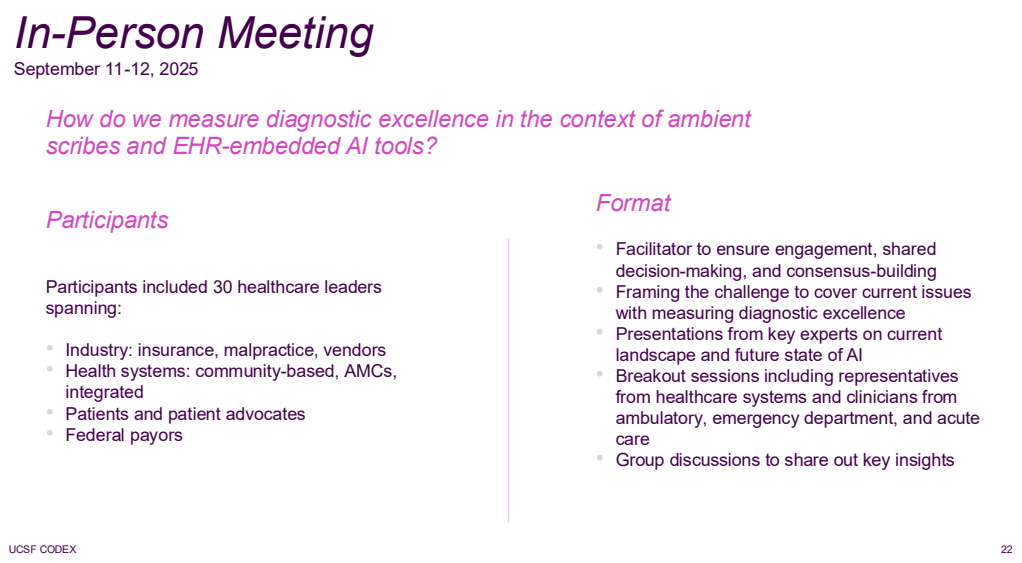


ROSNER: So we spent a good bit of time, as you can see here in advance of the Action Incubator developing the model with the idea then of bringing all of these stakeholders together in an in-person meeting. We held this first one this past September, and that in-person meeting really sort of sets the stage for quite a bit of homework to do afterwards. We're in that process right now which is to develop this actionable measurement framework.

I will say one of the interesting challenges in thinking about that when you are trying to develop something that can represent a diversity of interests and stakeholders is how to approach something that will generalize across many different healthcare settings and across many different stakeholders in a way that will be meaningful. And then at the conclusion of developing this framework, then the idea is to bring that framework out into the real

world — into health systems — and measure the actual impact for diagnostic excellence (DxEx).

[SLIDE 22]



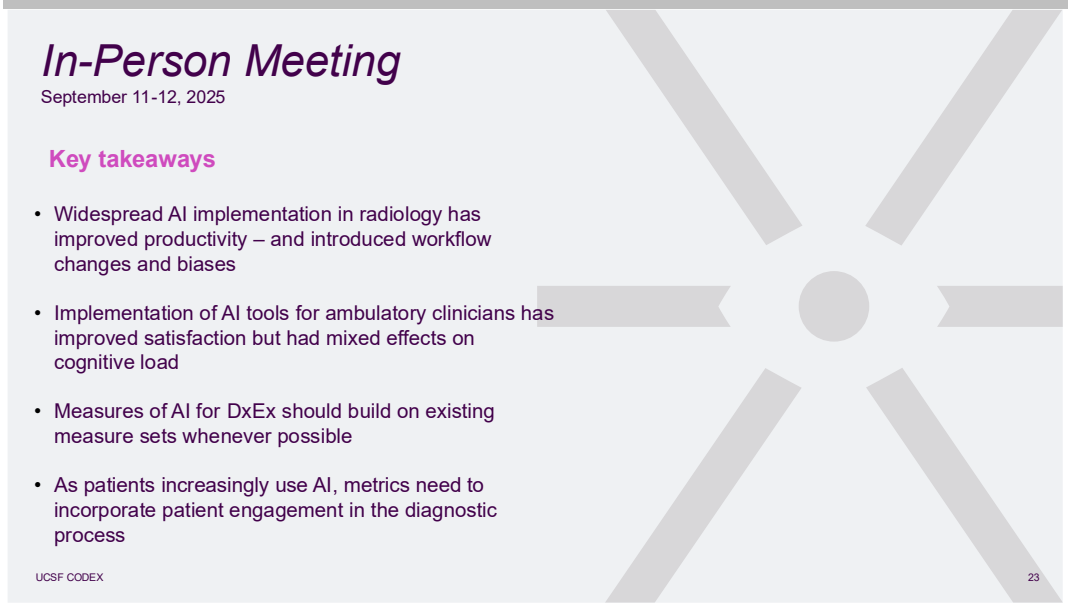
ROSNER: So we wanted to bring to the table the concept of how do we measure diagnostic excellence (DxEx) in the context of artificial intelligence (AI). Now I mentioned first a diversity of stakeholders, so we had about 30 or so stakeholders from across the county that represented really a broad landscape of participants. These included members from industry. We actually had folks from the AI scribe industry present from insurance, medical malpractice. We had representatives from eleven (11) health systems participating. The health systems are actually where this deployment would happen. So you can imagine the Action Incubator being the source of idea generation and collaboration, with an ultimate goal of deploying something to be measured across these health systems.

Now I'll say that it was important to us that these health systems represented as broad a swath as possible. So we had a split between academic medical centers, as well as community health systems, as well as safety net hospitals in order to try to capture that breadth of experience. We also made sure that we had patient advocacy organizations present at the table, as well as a couple of payors, one of which was a federal payor.

The way that we brought this group together was to really sit down in the same room together, but we had a facilitator organizing the discussion, sort of setting the ground rules. You can imagine when you put 30 or so experts into a room together, you may end up with 40 opinions and lots of people eager to participate. Fortunately, there was a lot of eagerness to participate and a lot of great ideas that came to the table, but the facilitator sort of helped manage all that, so that we would have some very discrete and actionable outcomes at the end of this in-person meeting that sort of set the blueprint for how to move forward in the months that followed.

We started the Action Incubator with a presentation from an expert in AI and radiology, because we felt that there are a lot of lessons that could be learned with radiology's experience with AI, since it's been a leader in that area, ahead of many other specialties. So we had expert presentations, but we also had breakout sessions where representatives sort of dove in more deeply into specific use cases and settings, including how do we think about diagnostic excellence (DxEx) using AI in ambulatory, emergency department (ED), acute care settings, et cetera. And then we had those groups share out some of their approaches and learnings.

[SLIDE 23]



ROSNER: Now one of the interesting challenges as you think about well, how is AI available or deployed in a very broad setting? Right now, as Sumant mentioned, the most broad use case of AI in healthcare is actually AI scribes. Now traditionally AI scribes are not meant for diagnostic purposes. They're meant for documentation, but because they're widely deployed we thought it might be interesting to look at how is the use of AI scribes potentially influencing diagnostic excellence (DxEx). We refer to this as the "current state" of AI deployment.

Now, in the interim since we had our Action Incubator meeting, things are moving towards what we refer to as the "future state" of AI implementation. That future state is one in which AI is deeply embedded into the electronic health record (EHR) so that clinicians can have, for example, conversational queries that are context-aware of the patient. You can imagine in this future state which is emerging and some health

systems have this now where clinicians might ask for differential diagnoses in a context-aware manner in which the artificial intelligence (AI) knows a lot about the patient because it's able to access the full medical record. But in the current state in the use of ambient AI scribes, one of the things that we posited was "well, does the use of ambient AI scribes in any way have an impact on cognitive load or cognitive burden?"

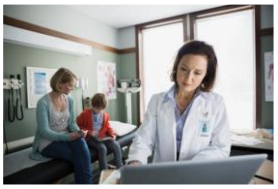
In the early days of this Action Incubator the literature had not yet sort of landed on an answer to that, but it does appear increasingly that there is a reduction in cognitive load when clinicians are using AI scribes. So the question is "does that give them cognitive space to be better diagnosticians? And, if so, how should we measure that?"

So we know that with this increasing adoption of AI scribes, increasingly clinicians are using them. Also, there's the potential for patients then to have input on their thoughts as to their diagnostic excellence (DxEx) when the clinician may be looking at them eye to eye in the encounter, rather than tethered to a keyboard.

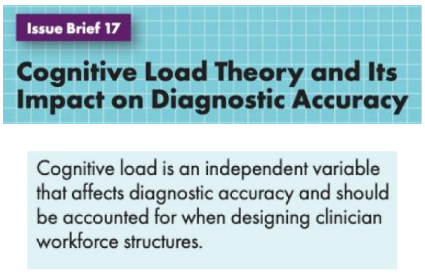
[SLIDE 24]

Evaluation in the Context of Current State of Widely Deployed AI

Traditional Care

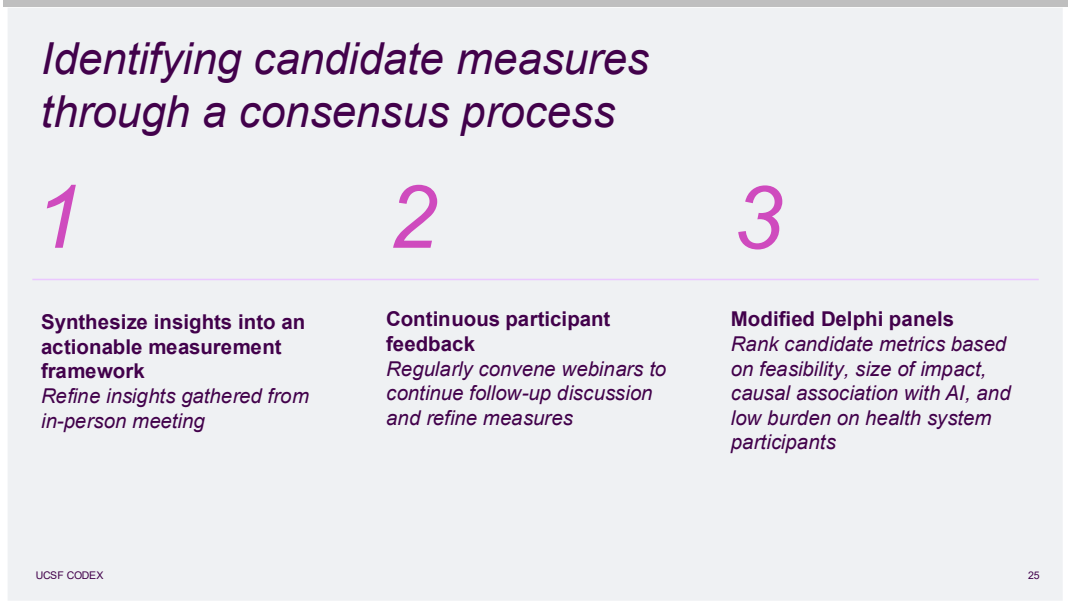


AI Scribe



ROSNER: So there's an interesting AHRQ issue brief which in fact supports this notion that cognitive load itself is an independent variable that can affect diagnostic excellence (DxEx) and should be accounted for when designing clinician workforce structures. So the way that our Action Incubator thought about landing on this was let's look at some kind of measures of diagnostic excellence in two contexts — one in traditional care, the image on the left where the clinician is in fact at that keyboard and documenting during the encounter vs. the one on the right where the clinician is freed from a keyboard using an AI scribe to do documentation.

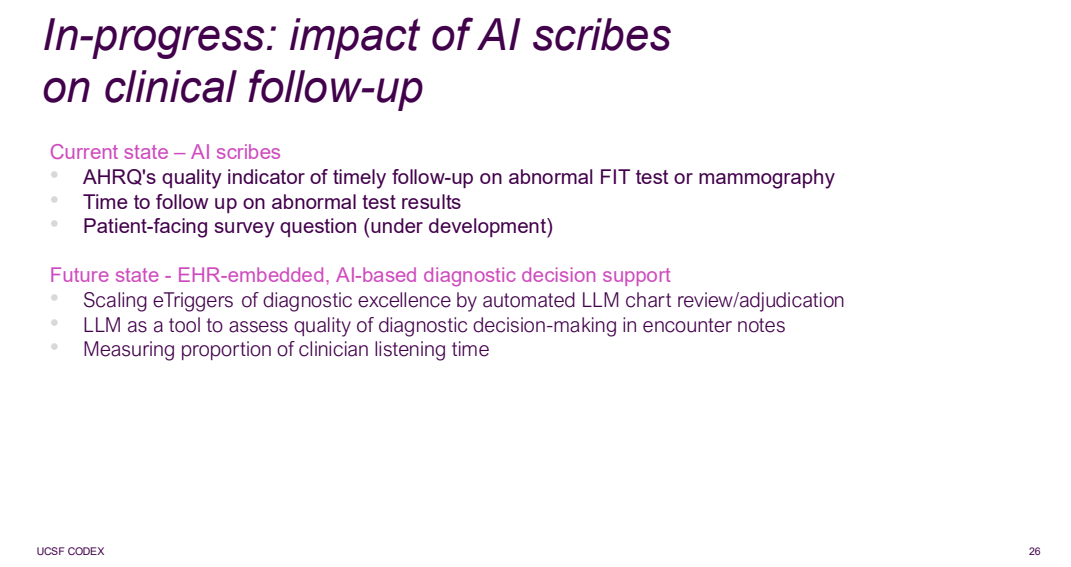
[SLIDE 25]



ROSNER: And so then our group turned to "well, what is it that we ought to measure?" This discussion really in essence was a modified Delphi panel where we proposed a variety of potential candidate metrics to measure. And then as a group we would vote on them and try to come to consensus on what would make the most sense. So since the in-person meeting we have hosted now a number of synchronous webinars or remote meetings where we discussed candidate metrics. The idea behind the metrics is to potentially have as broad an impact as possible, but in certain important areas. Now one of the challenges that Sumant has mentioned is that it's very difficult to have a purely generalizable measure of diagnostic excellence (DxEx). In fact, we find ourselves because of the variability of diagnosis condition or disease by disease, we find ourselves in fact gravitationally pulled towards thinking about how to measure diagnostic excellence more on a disease-by-disease basis. But if we can select diseases that have large numbers of patients and very meaningful

potential outcomes, that could be then very important as a proof of concept in moving this forward.

[SLIDE 26]



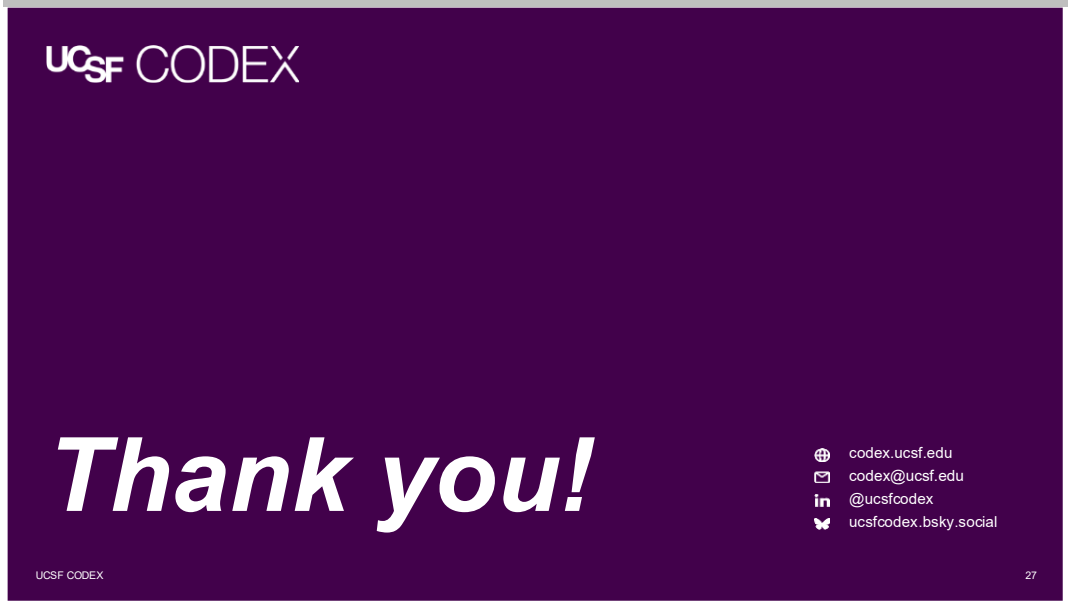
ROSNER: So where we have landed with our Action Incubator participants was the notion of looking at follow-up on abnormal test results. In particular, we landed on AHRQ's Quality Indicators (QI) for timely follow-up on abnormal stool tests for colorectal cancer and abnormal mammography tests for the detection of breast cancer. We chose these in particular because they're very well-defined actions that can be measured on the backend in the electronic health record (EHR). Did a clinician take the appropriate action? For example, following an abnormal screening FIT Test did they order a colonoscopy? That can readily be queried for in the electronic health record (EHR), and so it's easy to measure if that action was taken, and the same approach for abnormal mammograms.

So after landing on those two, because they do actually have substantial impact on a large population, we also wanted to make sure that we were capturing the patient experience of diagnostic excellence (DxEx) and what their perception of diagnostic excellence might be in the context of when a clinician uses a scribe vs. when they do not. We're still in the process of identifying what are appropriate questions to ask patients of that diagnostic experience.

Last, as I mentioned, we are looking towards this future state when AI is deeply embedded into the EHR. There are a number of health systems in which this is already available, but realistically in the next year we will see this much more broadly available where most of our participating health systems will have access to these sorts of tools — not only because vendors are developing them, but because also EHR vendors themselves are increasingly making these tools available in the electronic health record (EHR).

And then we will be able to look at a whole host of very interesting opportunities to measure diagnostic excellence (DxEx) looking at, you know, in what are e-TRIGGERs which have traditionally required human adjudication, but could LLMs do some of that adjudication for e-TRIGGERs looking at whether LLMs can be used to assess the quality of diagnostic decision-making, or even something as simple as purely just measuring the proportion of time that the clinician spends listening to the patient, which we know has an important effect on their potential performance as a diagnostician.

[SLIDE 27]



ROSNER: So with that, that wraps up the Action Incubator discussion and most of our presentation here. Sumant, if you would like to say anything to wrap up, feel free; otherwise, I think at this point we are interested in opening up the floor for Q&A.

RANJI: Thank you, Ben. Again, thanks to CMS and Battelle for giving us the opportunity to present our work. So hopefully what all of you have had a chance to understand from us is how we at CODEX are thinking about the measurement of diagnostic excellence (DxEx) in the era of artificial intelligence (AI). We think that artificial intelligence (AI), as it's being deployed more broadly within the healthcare system, has great potential for improving diagnosis along the dimensions we discussed — improving the accuracy of diagnosis, improving the timeliness of diagnosis, and ideally improving communication with patients so that they can understand diagnoses and the implications thereof.

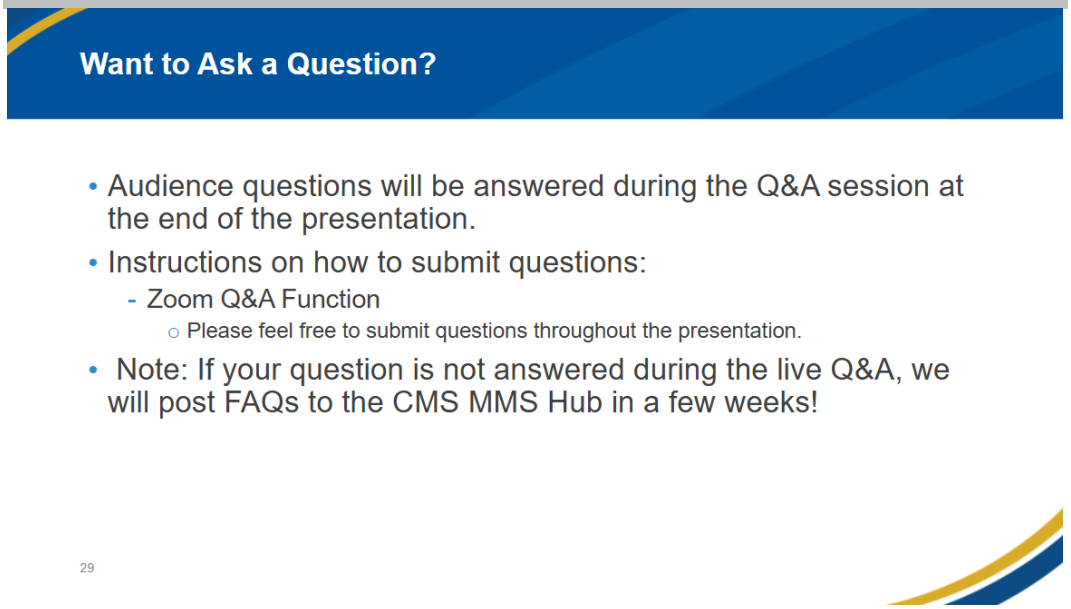
But then there are a lot of unanswered questions, and our Action Incubator is really our first step towards trying to answer some of those questions — particularly around as health systems are being asked to make large investments in AI-based applications, how will they know that these applications are actually reaching their potential in terms of improving diagnosis.

So as we are grappling with those questions, as Ben said, we decided that the best way to approach this was not necessarily to create new measures, but to think about the ways that AI could build on what's already been done in the measurement world and actually improve performance on some of the validated and endorsed measures already out there. So this, as Ben said, is very much a work in progress. We actually had our meeting earlier this week to try and finalize these measures, and so we would very much appreciate any thoughts or feedback. We are happy to take questions as well.

[SLIDE 28]

Q&A Portion Not Included in Transcript

[SLIDE 29]



Q&A Portion Not Included in Transcript

[SLIDE 30]

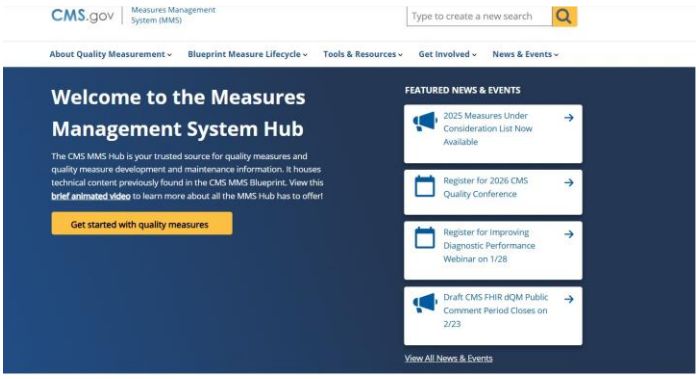
https://www.cmsqualcon.com/

30

GHUNNEY: So as we close out, we just want to share that CMS will be hosting their annual quality conference March 16-18th in Baltimore, Maryland. The theme for this year's session is: *Make America Healthy Again: Innovating Together for Better Health.* We encourage you to visit cmsqualcon.com to register for this wonderful opportunity.

[SLIDE 31]

Get Connected



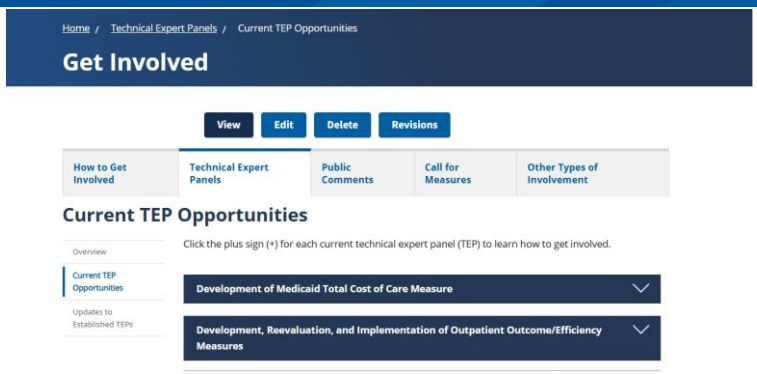
Visit: https://mmshub.cms.gov/

31

GHUNNEY: And as always, we encourage you to visit the MMS Hub which is your trusted source for quality measures and quality measure development and maintenance information. You can find all updates, and news and events that are happening in the quality measurement community such as if you go there right now, you'll see that there's currently a public comment period open for the "Draft CMS FHIR dQMs," which closes on February 23rd.

[SLIDE 32]

Get Involved



https://mmshub.cms.gov/get-involved/technical-expert-panel/current

32

GHUNNEY: Lastly, we also keep you updated on ways to get involved. There are currently open opportunities available for you to nominate yourself or others to technical expert panels (TEPs) for two measures. So visit the MMS Hub website for more information on how to get involved in those opportunities. And with that, I want to thank you again for taking the time to join us today. We look forward to speaking with you again soon.

[SLIDE 33]

GHUNNEY: If you have any questions, feel free to reach out to us at MMSsupport@battelle.org, and we would be happy to assist. Thank you again for coming today.

WEBINAR CONCLUDES