UNITED STATES OF AMERICA

FOOD AND DRUG ADMINISTRATION

+++

CENTER FOR DEVICES & RADIOLOGICAL HEALTH

+++

DIGITAL HEALTH ADVISORY COMMITTEE

DAY 2 CONFERENCE

+++

TOTAL PRODUCT LIFECYCLE CONSIDERATIONS FOR GENERATIVE AI-ENABLED DEVICES

+++

November 21, 2024

9:00 a.m. EST

On-Site Conference

Transcript Produced By:

ACSI Translations

1025 Connecticut Avenue, NW, Suite 1000,

Washington, DC 20036

https://acsitranslations.com/

Participants

Chairperson	Ami Bhatt, M.D.	Chief Innovation Officer (CIO) at the American College of Cardiology	Newton MA
Industry Representative	Diana Miller	Senior Director, Data Science Medtronic	Northridge, CA
Consumer Representative	Melissa Clarkson, Ph.D.	Assistant Professor of Biomedical Informatics, University of Kentucky College of Medicine	Lexington, KY
Voting Member	Jessica Jackson, Ph.D.	Founder and CEO Therapy Is For Everyone Psychological & Consultation Services, PLLC	Houston, TX
Voting Member	Thomas Maddox, M.D., MSc.	Vice President, Digital Products and Innovation, at BJC HealthCare, and Professor of Medicine (Cardiology), Washington University School of Medicine	St. Louis, MO
Voting Member	Chevon Rariy, M.D.	Chief Health Officer, and SVP Digital Health, Oncology Care Partners	Mequon, WI
Voting Member	Laura Stanley, Ph.D., CPE	Associate Professor, Gianforte School of Computing, and Director of Human Interaction Lab Montana State University, Bozeman, MT, and Clinical Associate Professor, Clemson University School of Health Research	Clemson, SC

Temporary Non-Voting Member	Taxiarchis Botsis, MSc., M.P.S., Ph.D.	Associate Professor of Oncology and Medicine The Sidney Kimmel Comprehensive Cancer Center Johns Hopkins University School of Medicine	Baltimore, MD
Temporary Non-Voting Member	Peter Elkin, M.D., MACP, FACMI, FNYAM, FAMIA, FIAHSI	UB Distinguished Professor and Chair Department of Biomedical Informatics Professor of Pathology and Anatomical Sciences University at Buffalo State University of New York	Buffalo, NY
Temporary Non-Voting Member	Rita Kukafka, DrPH, MA, FACMI	Professor of Biomedical Informatics and Sociomedical Sciences Department of Biomedical Informatics Chief Diversity Officer Columbia University	New York, NY
Temporary Non-Voting Member	Steven Posnack, MIS, MS	Deputy National Coordinator for Health IT in the Office of the National Coordinator for Health Information Technology, US DHHS	Washington, DC
Temporary Non-Voting Member	Pratik Shah, Ph.D.	Assistant Professor, Pathology University of California	Irvine, CA
Temporary Non-Voting Member	Apurv Soni, M.D., Ph.D.	Assistant Professor of Medicine and Director of Program in Digital Medicine Health System Science Center for Digital Health Solutions, Department of	Worcester, MA

		Medicine UMass Chan Medical School	
Temporary Non-Voting Member	Jagdish Khubchandani, MBBS, PhD, MPH	Professor of Public Health at New Mexico State University	Las Cruces, NM
Temporary Non-Voting Member	Thomas Radman, Ph.D.	Program Director Digital & Mobile Technologies Section, Division of Clinical Innovation National Center for Advancing Translational Sciences, National Institutes of Health, DHHS	Bethesda, MD
US Food and Drug Administration	Troy Tazbaz	Director of the Digital Health Center of Excellence, US Food and Drug Administration	Silver Spring, MD
US Food and Drug Administration	Sonja Fulmer, Ph.D.	Deputy Director, Digital Health Center of Excellence, US Food and Drug Administration	Silver Spring, MD
US Food and Drug Administration	Matthew Diamond, MD, Ph.D.	Chief Medical Officer, Digital Health Center of Excellence, US Food and Drug Administration	Silver Spring, MD
US Food and Drug Administration	James P. Swink	Designated Federal Officer US Food and Drug Administration	Silver Spring, MD

Table of Contents

Digital Health Advisory Committee (DHAC) Call to Order

Dr. Bhatt: Good morning. Welcome to day two. If everyone can please have a seat. We're a little more sparse in the audience this morning. We can all turn our heads and look every time someone comes in late.

All right, I would like to call this meeting of the FDA's Digital Health Advisory Committee on November 21st, 2024 to order. I am Dr. Ami Bhatt, Chairperson of this Committee. I am a Cardiologist in training previously at Mass General in Harvard and now the Chief Innovation Officer at the American College of Cardiology and honored to have this position. I note for the record that the non-voting members constitute a quorum as required by 21 C.F.R., Part 14. I would also like to add that the Committee members participating in today's meeting have received training in FDA device law and regulations.

For today's agenda, the Committee will discuss and make recommendations on how the use of generative AI may impact safety and effectiveness of medical devices enabled with this technology. The Committee will discuss premarket performance evaluation, risk management, and postmarket performance monitoring for generative AI-enabled devices. I'm going to go off-script. Yesterday we did a great job with premarket performance evaluation risk management. Today's focus is postmarket monitoring.

Welcome to day two. I would like to ask our distinguished Committee members and FDA experts to introduce themselves. I will call the first name. Please state your area of expertise, position, and affiliation, and then pass it on to the next person. Matthew Diamond, may I start with you?

Dr. Diamond: Yes. Matthew Diamond, Chief Medical Officer at FDA's Digital Health Center of Excellence.

Dr. Fulmer: Hi, Sonja Fulmer, Deputy Director for the Digital Health Center of Excellence at FDA.

Mr. Tazbaz: Good morning. Troy Tazbaz, Director of Digital Health Center of Excellence at FDA.

Dr. Radman: I'm Thomas Radman. I'm a Program Director at the National Center for Advancing Translational Sciences at the National Institutes of Health. And what that means is I oversee a portfolio of grants that fund infrastructure and research in translational science to increase the efficiency of getting basic research into products we would use in our everyday lives. Thank you.

Dr. Khubchandani: Good morning. Jagdish Khubchandani, Professor of Public Health, New Mexico State University. Interest in Social Epidemiology. Thank you.

Dr. Soni: Hello. Apurv Soni. Program in Digital Medicine director at UMass Chan Medical School and our group focuses on how we use technology to bring research to patients' homes and clinical care to patients' homes.

Dr. Rariy: Good morning. I'm Chevon Rariy. I'm a Physician Technology Executive. I'm the Chief Health Officer at a venture-backed startup Oncology Care Partners. We sit at the intersection of care delivery, data, and analytics with an underlining of health equity, and value-based care. Very excited to be here today and thank you.

Dr. Maddox: Good morning, Tom Maddox. I'm a Cardiologist at WashU School of Medicine in St. Louis, and its Partner Health System, BJC HealthCare, and I lead the Healthcare Innovation Lab there.

Mr. Swink: James Swink. I'm the Designated Federal Officer and Team Lead at CDRH.

Dr. Jackson: Good morning, Jessica Jackson. I'm a Licensed Psychologist, and I'm an Advisor and Consultant for digital health startups.

Dr. Shah: Morning, I'm Pratik Shah. I'm a Faculty Member and Professor at University of California, and I lead a group that looks at clinical deployment of generated AI technologies.

Mr. Posnack: Steven Posnack from the HHS Office of Assistant Secretary for Technology Policy.

Dr. Kukafka: Rita Kukafka. I'm a Professor of Biomedical Informatics and Sociomedical Sciences at Columbia University.

Dr. Elkin: Hello, I'm Peter Elkin. I am a distinguished Professor and Chair of the Department of Biomedical Informatics at the University of Buffalo, and my interest is in artificial intelligence and making the world a better place. Thank you very much.

Dr. Botsis: Good morning. I'm Taxiarchis Botsis, Associate Professor of Oncology Medicine at John Hopkins University and also Director of the Biomedical Informatics Corps in our Cancer Center.

Dr. Stanley: Good morning, everyone. I'm Laura Stanley. I'm a Computer Science Professor at Montana State University of Bozeman and my interests are in human AI integration, human systems integration. Thank you.

Dr. Clarkson: Hello, I'm Melissa Clarkson. I'm the Consumer Representative for the Committee. I'm also an Assistant Professor in Biomedical Informatics at the University of Kentucky.

Ms. Miller: Good morning. Diana Miller, Senior Director, Data Science Medtronic. I'm here as the Industry Representative. Thank you.

Dr. Bhatt: Great. Thank you all. James Swink, the Designated Federal Officer for the Digital Health Advisory Committee will make some introductory remarks. James?

Conflict of Interest Statement

Mr. Swink: Great. Bear with me as I read the conflict-of-interest statement. The Food and Drug Administration is convening today's meeting of the Digital Health Advisory Committee under the authority of the Federal Advisory Committee Act of 1972. With the exception of the industry representative, all members and consultants of the Committee are special government employees or regular federal employees from other agencies and are subject to federal conflict of interest laws and regulations.

The following information on the status of this Committee's compliance with federal ethics and conflict of interest laws is covered by, but not limited to, those found at 18 U.S.C. §208, are being provided to participants in today's meeting and to the public.

FDA has determined that members and consultants of this Committee are in compliance with federal ethics and conflict of interest laws. Under 18 U.S.C. §208, Congress has authorized FDA to grant waivers to special government employees and regular federal employees who have financial conflicts when it is determined that the agency's need for a particular individual's services outweighs his or her potential financial conflict of interest.

Related to discussion of today's meetings members and consulted as Committee who or special government employees or regular federal employees have been screened for potential financial conflicts of interest of their own as well as those imputed to them, including those of their spouses or minor children and, for purposes of 18 U.S.C. §208, their employers. These interests may include investments, consulting, expert witness testimony, contracts, grants, CRADAs, creatives, teaching, speaking, writing, patents, royalties, and primary employment.

For today's agenda, the Committee will discuss how the use of generative artificial intelligence may impact safety and effectiveness of medical devices enabled with this technology. The Committee will discuss postmarket performance and based on the agenda for today's meeting in all financial interest reported by the Committee members and consultants, no conflict-of-interest waivers have been issued in accordance with 18 U.S.C. §208.

Ms. Diana Miller is serving as the Industry Representative for generative AI, the large language models, and is acting on behalf of all related industry. Ms. Miller is employed by Medtronic plc. For the record, the agency notes that Chris Longhurst and

Dr. Grace Cordovano, who are invited guest speakers with us today, have reported no interest in relation to today's meeting.

We would like to remind members and consultants that if the discussions involve any other products or firms not already on the agenda for which an FDA participant has a personal or imputed financial interest, the participants need to exclude themselves from such involvement and their exclusion will be noted for the record.

FDA encourages all other participants to advise the Panel of any financial relationship they may have with any firms at issue.

A copy of the statement will be available for review and will be included as a part of the official transcript.

Before I turn the meeting back to Dr. Bhatt, I would like to make a few general announcements. Transcripts of today's meeting will be available from ACSI. Their phone number is (202) 599-8456. Information on purchasing video of today's meeting and handouts are at the registration table. The FDA press contact for today's meeting is James McKinney. All written comments received provided to the Panel and the FDA review prior to today's meeting. There is an active docket on the FDA website, and it's open to January 21st. So, you may submit your dockets comments there.

I would like to remind everyone that members of the public and the press are not permitted in the Panel area, which is the area beyond the speaker's podium. If you're presenting in an Open Public Hearing Session and have not previously provided an electronic copy of your slide presentation, please arrange to do so with Mr. Artair Mallett at the registration table.

And in order to help the transcriptionist to identify who is speaking, please be sure to identify yourself each and every time that you speak. A reminder that lunch today will be a bit later today at 1:00 PM. And please silence your cell phones and other electronic devices at this time. Thank you.

Dr. Bhatt: Thank you so much. I want to provide a brief overview of how today's meeting will run. So, first we'll have a recap of yesterday by the FDA by Aubrey Shick. After that, we will have an Open Public Hearing, and then we'll have a chance to ask questions of those who present in the Open Public Hearing. We will then have presentations that are specifically oriented toward thinking about postmarket performance monitoring, which will then be the question after lunch that we will discuss.

After going through yesterday, we learned a lot of things after doing question one and two, and so I did want to ask Troy Tazbaz to just say a word or two about level setting for today and what the FDA needs from us. It's an inaugural Committee, and so I think it's good that we remind ourselves how we can be most helpful.

Mr. Tazbaz: Thank you Dr. Bhatt. Troy Tazbaz with the FDA. So, yesterday we were able to identify many problems with this transformative technology, and we've focused on the potential, but all of the issues that we have to address for this potential to be a reality. And throughout, you know, history in the United States, public-private partnerships have tackled very challenging and complex issues to actually push innovation forward. And this is one of those opportunities for us to collaborate as an entire ecosystem, and that's government, that's academia, that's the industry, to push this innovation forward in a way that is safe, effective, has the guardrails to be applied

to the criticality of the nature of healthcare. And so, yesterday we did a phenomenal job in identifying some of the challenges that we have. And we heard this from the public, we've heard this from some of the presenters, and we heard this throughout the conversations that we had throughout the Committee.

And so, what we have to do moving forward is start thinking about the solutions. How do we prioritize some of these, tackling these issues? And so, at the end of this, our goal, and when I say the end of this, I'm not talking about end of today, I'm talking about the end of this Advisory Committee, is that a set of priorities that we collectively have to tackle? And from our point of view that would be things like, well, what do we have to solve for? Do we need different evaluation frameworks? Do we need different guidance? How do we think about regulation when it comes to generative AI? And so, we're looking for those types of answers and the priorities.

And so, our goal from the government side is to get the input from you all around what should be our priorities, what should we exclude from that? And there are a ton of issues that we identified, but we have to prioritize those issues to say which ones are really the important ones that we have to solve for today and then tackle the other ones maybe down the line. And so, that's what I would like to get out of this. Thank you.

Dr. Bhatt: Thank you so much. So, I specifically was wanting Troy to speak about this because yesterday we did premarket and I think the nice thing about premarket is many of us are familiar and comfortable with what it takes to get the right technologies to the right place, and it was a robust discussion that actually ran over, but we did make up time at the end. Risk management was where we brought up a lot of questions,

limited in answers only because it requires really a collaborative effort across industries to get there.

But today, this is our opportunity for postmarket performance monitoring. We've done it before through the FDA for drugs, for devices, but generative AI is different. And so, as we're listening to our public speakers today in our open session as we're listening to our invited speakers, and then in the afternoon when we talk, the goal is really for us to create the infrastructure here amongst this group to start saying, here is how we can think about postmarket surveillance for something as challenging as generative AI, here are the most important parts of that infrastructure, and then here are the guardrails that come with it.

And so, just with that in mind today, I'm excited for us to get started. We will start with Aubrey Shick, Senior Digital Advisor for the Digital Health Center of Excellence in CDRH to give us a recap of day one.

Recap of Meeting Day 1

Ms. Shick: Thank you, Dr. Bhatt, and good morning again. I'm Aubrey Shick. I'd like to thank all of the invited speakers, Breakout Session participants, and Open Public Hearing Speakers for their thoughtful perspectives during day one of the meeting. We heard a few recurrent themes throughout yesterday's discussions, which focused on premarket performance evaluation and risk management. These themes included supporting trust, transparency, and usability, and taking a risk-based approach to the regulation of GenAI-enabled devices.

We began the day with Commissioner Califf who acknowledged the potential for generative AI to improve the quality and delivery of care in clinical settings and the

home. Next, CDRH's Director, Dr. Tarver, emphasized the need to keep diverse and underserved populations top of mind to provide equitable access to AI technologies that are high quality, safe, and effective, and to ultimately help move the needle on chronic diseases. Mr. Tazbaz, Director of the Digital Health Center of Excellence, highlighted the importance and timeliness of addressing generative AI-enabled devices and added the total product lifecycle approach to the discussion.

During the premarket performance evaluation section, the discussion included defining the scope of device intended use, the prominence of foundation models implemented in medical devices, the importance of risk commensurate oversight, the need to understand and limit hallucinations and their impact and that diverse data are both needed and sometimes challenging to find. We also heard about current work related to the performance measurement of GenAI.

The Committee discussed how a device description or characterization should include the device's intended use and specific use cases as well as cases that may introduce uncertainty, the intended population, including demographic information, where the device is intended to be used and what data the device was trained on, including the dataset size, types, and composition of patient population. We heard discussion on how a standard data sheet or model card may be a helpful way to provide this information.

The Committee also discussed how description or characterization may include information about whether a device enabled by generative AI is adaptive or not, and what the intended role of the AI is in the clinical workflow. The Committee discussed how the device's premarket evaluation could include a characterization of the device's

performance, including performance in different populations and settings as applicable for the device's intended use. The Committee discussed how premarket evaluation could characterize repeatability, reproducibility and uncertainty, including uncertainty estimates, hallucination rates, or error rates. They noted that premarket evaluation could include a proposal for postmarket monitoring. They discussed that the types and level of information should generally be commensurate with device risk, consistent with FDA's existing risk-based approach. The Committee also posited stress testing as a method of risk mitigation.

With respect to the user interface, the Committee discussed how an explanation of the device inputs and outputs is an important consideration in ensuring transparency to the user for devices enabled by generative AI. The Committee further discussed the topic of transparency, noting that it may be important for users to know that an output was created by a device enabled with generative AI and that such outputs may not be able to be consistently reproduced. The Committee discussed the need to communicate device performance to FDA and users. This includes sensitivity and specificity for the specific intended use of the device. The Committee recognized that data drift may impact safety and performance of the models.

During the risk management section, the Committee discussed strategies, controls, governance, and frameworks to mitigate risk of generative AI applications as well as risk related to ethics and usability. For example, we heard about the importance of user training. We also heard that these mitigations could include a focus on transparency, explainability, and real-world performance.

The Committee addressed the FDA's questions about new opportunities that have been enabled by generative AI for medical devices and new controls that may be needed to mitigate risks associated with generative AI technologies that enable those opportunities. GenAI can provide new ways of presenting information that may seem more human-like and give the impression of human intelligence to users. The Committee discussed how digital health literacy is essential to consider for patients, particularly in consideration of health equity.

Additionally, the Committee noted that the risk of patient harm is a central consideration for risk management and governance. For example, governance may need to differ across different types and implementations of generative AI to leverage site specific or global governance to best protect patients. The Committee considered how the responsibility for generative AI-enabled devices may ultimately be shared among manufacturers, physicians, and the health systems deploying them.

We look forward to hearing today's presentations as well as the Committee's deliberations on questions. And with that, I turn it over to you, Dr. Bhatt.

Open Public Hearing

Dr. Bhatt: Thank you so much, Ms. Shick. We will now proceed with the Open Public Hearing of this advisory Panel meeting. For the record, all Panel members have been provided written comments received prior to this meeting for their consideration.

During the Open Public Hearing, public attendees are given an opportunity to address the Panel to present data, information, or views relevant to the meeting agenda. Mr. Swink will now read the Open Public Hearing disclosure process statement.

Mr. Swink: Thank you. Both the Food and Drug Administration and the public believe in a transparent process for information gathering and decision-making. To ensure such transparency at the Open Public Hearing Session of the Public Advisory Panel, FDA believes that it is important to understand the context of individual state presentation. For this reason, FDA encourages you, the Open Public Hearing speaker, at the beginning of your written or oral statement to advise the Committee of any financial relationships that you may have with any company or group that may be affected by the topic of this meeting. For example, this financial information may include accompanies or group's payment of your travel, lodging, or other expenses in connection with your attendance at this meeting. Likewise, FDA encourages you at the beginning of your statement to revise the Committee if you do have any such financial relationships. If you choose not to address this issue of financial relationships at the beginning of your statement, it will not preclude you from speaking.

Dr. Bhatt: Thank you. The FDA and this Panel place great importance in the Open Public Hearing process. The insights and comments provided can help the agency and this Panel in their consideration of the issues before them. We ask that each presenter speak clearly to allow the transcriptionist— You have to love when you can't say transcriptionist after you say, "speak clearly." —to provide an accurate transcription of the proceedings of this meeting.

The Panel appreciates that each speaker remains cognizant of their speaking time. I will today watch closely. I have a good bird's eye view now of the green, yellow, and red. So, at yellow, you'll notice me actually just raising my hand at you, that way I'm not speaking and interrupting you. And at red, if we could stay on time, that would be great. There is a lot to get through today, and we're excited for it.

So, we have seven requests to speak. The first two will be virtual presentations. So, we can start with the video from Dr. Robert Steinbrook, Health Research Group Director of Public Citizen.

Dr. Steinbrook: I am Robert Steinbrook, Director of Public Citizen Health Research Group. We have no financial conflicts of interest. Public Citizen welcomes the FDA's comprehensive consideration of generative artificial intelligence-enabled devices. Public Citizen urges that disclosure and transparency to patients and healthcare professionals be required when GenAI is used in healthcare settings. We also urge that databases used for training such devices reflect the patient population that they are intended to serve to prevent discrimination and reduce bias.

For the protection of consumers and patients, enhanced scrutiny of health-related GenAI devices is essential. We're particularly concerned about the influences of big money and greed in our healthcare system. When companies cut corners and rapidly developing and implementing GenAI devices, patients are at risk for harm.

My comments address two topics. First, the need for additional regulatory oversight that is specific to GenAI-enabled devices to provide reasonable assurance of their safety and effectiveness, including new requirements for postmarket monitoring of device safety and performance. And second, the critical importance of detailed and transparent information for the individuals and medical professionals who use GenAI-enabled devices. Special scrutiny for GenAI-enabled devices, including consumer health-related tools and applications requires either presumptively designated these devices as Class III devices requiring premarket FDA approval for safety and efficacy, including compliance with the Department of Health and Human Services standards for

trustworthy artificial intelligence, or the establishment of a new and more stringent premarket approval system for such devices that are not designated as Class III.

In the executive summary for this meeting, the FDA states that it, quote, "may require special controls unique to GenAI-enabled devices when needed to provide reasonable assurance of the safety and effectiveness of the device", close quote, including certain Class II devices. We urge the agency to require such controls and the requirement to include robust postmarket monitoring of device safety and performance, notification requirements to the users of GenAI-enabled devices if the device safety and performance is not as attended, and procedures to promptly remove devices with newly recognized safety and performance concerns from the market.

The FDA should establish appropriate thresholds for suspending companies found to have repeatedly violated its rules and requirements and referring wrongdoing to the Department of Justice for legal action. Companies found to have knowingly concealed harms or substantial potential harms should face criminal prosecution for the company as well as top level responsible corporate officers.

The FDA has made constructive suggestions for the provision of transparent information about GenAI-enabled devices. For individuals, the critical point is that the GenAI-enabled device must not be a proprietary black box. To the maximum extent possible, such devices should be explained in easy-to-understand terms about their design, their autonomy, and the extent of autonomy. Complete information should be provided about the safety features to prevent hallucinations and other potentially dangerous, erroneous or false content, the anticipated frequency of such hallucinations

or false content, how people interact with the device, and the safety and other controls that the user may have.

Although GenAI-enabled devices have promised to improve health, the risks are not fully known. There is justified concern that without robust oversight, that is, oversight that is more stringent and demanding than current requirements, the risks of harm will substantially increase. We urge the FDA to move forward with strong and comprehensive regulatory requirements for GenAI-enabled devices.

I thank Eagan Kemp at Public Citizen for his help in preparing these comments. Thank you for the opportunity to comment.

Dr. Bhatt: Wonderful. Next, we have a video from Frederick Chen from the American Medical Association.

Dr. Chen: Good afternoon. My name is Dr. Frederick Chen, Chief Health and Science Officer at the American Medical Association. I appreciate the opportunity to provide public comment on the vital topic of augmented intelligence in healthcare. I'd like to offer the perspectives of practicing physicians like family doctors, dermatologists, and gastroenterologists who work in a variety of settings from rural to urban. These physicians may not be early adopters, but they see AI's potential and deserve to have their needs and concerns represented.

AI has transformative potential in medicine, from enhancing diagnostic precision to easing administrative workflows. Yet with this promise comes a set of crucial challenges. In December 2023, the AMA surveyed over a thousand physicians in different specialties to assess their views on AI. While 38% reported using some form

of AI in their practice, about two thirds remain cautious. Major concerns include data privacy, liability for AI errors and potential impacts on the doctor-patient relationship.

Today, I'm going to discuss four areas essential to the responsible integration of AI and healthcare: transparency, clinical validation, integration of AI into clinical workflows, and the importance of data privacy and security.

First, transparency is foundational to the ethical use of AI in healthcare. Physicians and patients alike need to understand when AI is part of a care process, especially for decisions that impact diagnosis or treatment plans. For transparency to be meaningful, AI developers have to share detailed information about how models are trained, what data sources are used, and any known limitations or biases. For example, a diagnostic AI tool might be accurate in a controlled setting, but it'll work really differently in a community hospital with a diverse patient population. While we were pleased to see FDA's release of guiding principles on transparency, the AMA is disappointed that we have not yet seen transparency mandates for AI-enabled medical devices and other healthcare AI. We strongly encourage FDA to consider what it can do, including updating labeling guidelines to further the necessary transparency for AI-enabled medical devices.

Second, robust validation of AI systems is essential to ensuring patient safety and clinical efficacy. AI models must undergo continuous testing and refinement to ensure that they are accurate, unbiased, and effective across a diverse range of patient populations and settings. As we know, models are often built on data that underrepresent certain populations, leading to potential inaccuracies that could harm those who are most vulnerable. We encourage the FDA to adopt rigorous and ongoing

validation requirements to ensure that AI tools perform safely and equitably, especially in real-world healthcare environments. The AMA strongly supports a risk-based approach to AI validation, where higher risk tools like those involved in diagnostics or treatment undergo more stringent testing than lower risk administrative tools. Additionally, the continued value of AI-enabled medical devices will depend largely on their safety and performance in the real world. Significant attention should be paid towards building post marketing surveillance systems tailored to AI-enabled medical devices.

Third, AI should augment rather than replace clinical judgment and include safeguards that allow for physician autonomy. While AI offers significant decision support capabilities, physicians must remain central in interpreting AI recommendations, especially when the stakes for patient outcomes are high. The AMA believes that human oversight is essential to prevent over-reliance on AI and to safeguard patient safety. For AI to genuinely support healthcare teams, it must fit within existing workflows in a way that enhances rather than disrupts the care process. AI tools should be designed with input from practicing clinicians to ensure that they align with real-world workflows and address the actual needs within the healthcare environment.

Finally, data privacy and security are paramount as we expand the use of AI in healthcare. AI models rely on vast amounts of data to function effectively, which means that patient data privacy must be safeguarded. AI tools can also be tempting targets for cyberattacks. Healthcare data, often sensitive and personal, require strict protections to ensure that patient information is not misused, re-identified or otherwise compromised. Patients and physicians should have rights to know how their data are used and even to opt out if they have concerns about privacy. As large language models and other

generative AI systems continue to develop, there's a risk that patient data might be inadvertently included in broader datasets.

The AMA envisions a future where AI serves as a valuable tool that enhances healthcare delivery without compromising patient safety, data integrity or physician autonomy. The AMA stands ready to work with the FDA to develop policies that ensure AI and healthcare is deployed responsibly with a steadfast commitment to patient safety, clinical efficacy and equity. Thank you for your time, and we look forward to continued collaboration on this critical issue.

Dr. Bhatt: Thank you to both of our videoed speakers. Next, we'll have two speakers who have five minutes each from Deloitte Consulting. We will start with Dr. Sandhya Polu, and then Dr. Anil Bhatta. Dr. Polu.

Dr. Polu: We're going to reverse that order.

Dr. Bhatt: Okay.

Dr. Bhatta: Good morning. We want to thank the Advisory Committee and the FDA for this opportunity to discuss some of the considerations around the use of safe and effective use of generative AI in medical devices. Should go to the next slide. Try to do this. It's just the right, right arrow? And the title of our presentation today is addressing the multi-use nature of generative AI in medical devices. So, my name is Anil Bhatta, I'm a Manager at Deloitte Consulting. I have a background in pharmacology, so I'm a Pharmacologist by training, and I'm a Biomedical Data Scientist by passion, and I've been working with FDA on a variety of different regulatory science projects as well as using AI and generative AI to make the regulatory review process more efficient.

So, today we want to talk about some of the unique challenges to generative AI and then how you go about addressing some of these challenges. So, generative AI has a lot of potential, and it can be transformative in the healthcare setting. However, because of the very dynamic and complex nature of generative AI models, there are various different challenges that we need to consider from a monitoring perspective, especially in a postmarketing setting. And a lot of challenges with traditional AI also apply to generative AI. However, there are several other challenges that are truly unique to generative AI, and therefore, you know, there's a need for a new framework in terms of addressing those challenges. So, here we have outlined five unique challenges with generative AI and then how you go about addressing those.

So, the first one centers around just the variability in unlimited output from a generative AI model. And because of this variability and the output, it becomes very challenging when you think through the quality of those outcomes. Unlike traditional AI where a definitive outcome, generative AI has variability and its outcome, therefore makes it very challenging from a quality perspective. And it's very paramount in a medical context where the accuracy and reliability of those outcomes are paramount.

There's also a challenge in terms of defining what an error is for generative AI. So, for a lot of generative AI models, unlike traditional AI where you're measuring the output of your model against a defined clear truth, for generative AI, the outputs aren't so straightforward, right? So, therefore, it's really hard to define exactly what error is. And because of those difficulties in defining error, it trickles down into from a model evaluation perspective, how do you develop metrics that can truly measure the performance of these generative AI models in these medical devices? So, unlike traditional AI where you can rely on some of the traditional metrics like accuracies and

recall and precision, even F1 scores, generative AI models, there are a lot of nuances where some of these metrics may not apply. So, there's this need for a new metric framework to evaluate these models.

And then there's also uncertainties around the unlimited amount of input that can feed into a generative AI, which can result in an unlimited amount of output. So, therefore there's complexity in monitoring that as well. And then lastly, it centers around resources that are needed to monitor these generative AI-based models in medical devices, in a postmarketing setting where there's a significant need for computational power for you from a monitoring perspective, right? Because a lot of these models are based on large language models with billions of parameters, which requires significant computational resources for monitoring. And subsequently, there is also significant need for human oversight because of a lack of standard model metrics, you're having to rely on a lot of human in the loop to validate these models.

So, because of these truly unique and novel challenges with generative AI, there are several risks that we want to present to you today. These risks we've bucketed for today's presentation into three main categories.

The first category, it centers around misinterpretation, and misinterpretation has several different layers. So, there is a misinterpretation at the model level. So, if you're injecting a prompt into a medical device, how it interacts with the underlying data in order for it to generate a response to your prompt? So, there's interpretation of the data at the model level. And then there's also interpretation, especially when we are talking about GenAI-enabled medical devices that are deployed in a nonclinical setting. So,

these are devices that patients take with them at home. How do they interpret the results of those generative AI devices?

The second risk category we want to discuss today is jailbreaking. So, this really centers around prompt injection or other types of malicious attacks to the generative AI systems, therefore generating output that are outside of its intended use. And it's really important when it comes to medical devices where we really want to make sure that we safeguard the use of generative AI, almost like a narrow use of generative AI in those contexts.

And then lastly, there are biases and overreliances and just like misinterpretation biases, there are different stakeholders in that ecosystem that you need to think about, you know, biases from a physician's or caregiver's perspective, how they interpret the outcome of the generative AI model from a patient perspective, you know, how do they interpret the outcome of those outputs, and then potential for misdiagnosis and potentially mis-self-treatment.

And then also there is the overreliance issue, and this is especially for physicians and other clinicians where they might tend to over rely on the outcome of the generative AI and not thinking through their own clinical judgment, right?

So, these are the risks and some of the challenges that we discussed, and I'm going to turn it to my colleague, Sandhya, to talk about some of the approaches that we have thought about for addressing these challenges and risks. Over to you, Sandhya.

Dr. Polu: Thank you. Good morning, everyone. I'm Sandhya Polu, and I'd like to focus a little bit on how we address some of the many challenges that were discussed yesterday, and that Anil just presented. At Deloitte, we have a trustworthy AI

framework with several dimensions that we use to support trustworthy AI outputs. It's similar to some of the frameworks across federal agencies, including HHS, and today I'd like to focus on five dimensions of that framework.

The first one is Safe and Secure, and how do we ensure that GenAI models produce safe and secure outputs? We can think about that from two different perspectives. One is a user perspective; one is a system perspective. So, from the user perspective, how do we ensure that user prompts, user inputs return useful, relevant, appropriate information? We can use prompt shields to constrain user prompts so that they are returning information that is appropriate to the context of use. We can also validate that user inputs are appropriate, for example, for patient facing devices that if let's say it's a continuous glucose monitor, they're not asking about some respiratory problem, for example. From the system perspective, Anil mentioned the risk of malicious attacks, prompt injections, which turn seemingly benign prompts into prompts that could return harmful, misleading false information. There are several methods to address prompt injection attacks. One is validating user inputs, for example, constraining the input length, sanitizing outputs to filter out any sensitive information. Also, building in safeguards to direct system prompts so that they only return information that is relevant to the context of use.

The second dimension is Robust and Reliable. This really is about ensuring that responses are contextually appropriate. Again, we can use system prompt engineering to support relevant and appropriate outputs. There are several techniques to minimize hallucinations as well, including domain specific fine-tuning.

The third dimension is Privacy, and here we really look from the patient perspective to focus on prompt safety so that sensitive PHI data is treated appropriately. We also want to create patient awareness, particularly for patient facing devices so that they're aware of how their data may or may not be used.

The fourth dimension is Transparent and Explainable. Trace back traceability to the source outputs are really important to ensure that transparency. Also, although we think about explainability in AI in terms of explaining the outputs developed by algorithms, we would like to propose thinking about explainability in terms of outputs that the patient may see. We can take inspiration from the proposed patient medication information rule, which would stipulate that patient medication information is an eighth-grade reading level, for example.

And the last dimension we'd like to talk about today is Fair and Impartial and really how do we manage bias in terms of GenAI? It's obviously not unique to GenAI, but the quantity of data sources is orders of magnitude greater, and we can address potential bias in AI throughout the GenAI lifecycle from designing and model development to deployment. Some of the techniques that are available or validating appropriate representation in the training datasets using synthetic data to do bias training datasets. We can also validate that the training dataset has the appropriate representation, pre-processed training data to filter out any bias language. And there's several other techniques available as well that are in development and research.

And lastly, we just want to emphasize the importance of human-centered design throughout the device product lifecycle and throughout the GenAI lifecycle and

ensuring that humans are always part of that process. Thank you, and we'll take any questions you have.

Dr. Bhatt: Great, thank you so much. We'll move on to the next speaker. We'll do questions at the end. Dr. Shrestha from Partner Innolitics, you have five minutes.

Dr. Shrestha: Hello, good morning. Yeah, so my name is Yujan Shrestha. It is very nice to be here. So, I'm affiliated with my company called Innolitics. They have paid for my hotel and flight, and I also owe my wife precisely 15 diaper changes as well. So, I want to also disclose that conflict of interest there. So, at my company Innolitics, we are a software development agency, and we also help our clients both build their software, medical devices, and AI medical devices and get them cleared through FDA. And so, today I would like to discuss a potential example of how I think we can turn postmarket surveillance from a stick into a carrot, where a stick is something that the manufacturers don't want to do, and a carrot is something that they would want to do. This is just one example, but I want to also explore how this can be done more generally as well.

So, a lot of my role as a consultant is to help my clients allocate their resources. They have a certain set of key marketing claims that they want to achieve, and they have a finite budget and timeline. If you know someone that doesn't have a finite budget timeline, please send them my way, I'll be sure to help them out. And so, everyone wants the same thing, patients, FDA industry, we all want the same thing. We want safe and effective medical devices quickly and with a low price point.

However, in the real world, trade-offs must be made because of budget and timeline constraints. And these trade-offs unfortunately mostly are made in the premarket, with the premarket clearance as that main focus. That's why most smaller-

stage startups, earlier-stage startups, are just trying to get to premarket clearance, and you can understand that they don't really care most about the postmarket at that time. And unfortunately, in the postmarket, this becomes a type of regulatory debt. And if you don't mind me using a software analogy, it's kind of like technical debt, where new features always take precedence over fixing old debt. And I believe the same thing happens to postmarket analysis there.

So, I believe that is one of the fundamental reasons why we only have a 4% update in radiology, as was discussed yesterday. And also, we have the issues of data drift and also lack of data generalization as well because of the lack of postmarket feedback.

So, let's change the narrative and turn it into a win-win situation for everybody. You know, why not turn postmarket evaluation into something that could be pitched to decrease premarket burden and to do so while also decreasing cost, time to market, reduce regional biases, mitigate data drift, and also mitigate model drift in the multi-layer model application design?

Here I would like to present a hypothetical device, and I promise I'm not getting free advice from you guys. This is purely hypothetical for now. This is an example of a device that has an open-ended input and a closed-ended output. It uses LLMs of unknown provenance, and what it does is it takes DICOM header files, and it tags them with certain key bullion classifiers for downstream things like hanging protocols and other standardization.

Here's an example of a synthetic data generation pipeline that uses site specific data and uses what I'm calling a frontier model. And I'm defining a frontier model as a

foundation model that is on the cutting edge. And that frontier model uses a specific test generation prompt or test data generation prompt that then creates synthetic data and then through some testing mechanisms that we talked about more yesterday by people smarter than me, we can verify this very focused intended use of this model. So, it's kind of like soup validation where the soup could do lots of things, but you only care about the things that you want, that you need that soup to do for your device. It's a similar concept here.

I also want to describe the concept of also doing validation for one more intended use. And that is to serve as a proxy for a ground truther for this very narrow task. And if we combine the two together, we have site-specific data, fully automated data synthetic generation, and we can even have a fully automated data ground truthing and also a nightly comparison to the device outputs. And you can imagine here, this operates nightly. Because it operates nightly it can detect data drift as it occurs. Also, the foundation models could be swapped out, so that can capture data drift in a more global level as well.

And just to conclude here, sorry about going a little bit over. So, I just want to say that I believe approaches like this would get the 90/10 trade-off balanced just right where we are giving something to the manufacturers that they can do in the premarket that would also help them in the postmarket as well. Thank you for your time.

Dr. Bhatt: Thank you very much. Next, we'll have Raaga Kanakam from Doctors for America. You'll have five minutes. Okay, we'll come back to them. Dr. Keith Dreyer, Chief Data Science Officer, Mass General Brigham. Dr. Dreyer. You have 10 minutes.

Dr. Dreyer: Pardon me?

Dr. Bhatt: 10 minutes.

Dr. Dreyer: Oh, I'll take it. Thank you. Hello everybody again, good morning. I want to thank the FDA for the creation of this ADCOM, it's truly been needed. I would just suggest that you're stepping into a movie that's been playing for seven years. The questions that you're asking about, say, specifically monitoring are questions we've been asking for seven years. I do know there are subtle nuances and additions and differences with regard to GenAI, foundation models, new technology. But as I sat and listened to all day yesterday, it's just a reminder of the last 10 years of what radiology has been trying to answer and deal with and petition and talk about, and address. So, I would just suggest, as the previous speaker said, there's technical debt here. What you are trying to resolve here with these discussions are not just the future, it's a thousand devices that exist today. There is no monitoring to these devices, and that's just not right. So, the notion of trying to solve this problem going forward, hopefully, you'll be thinking of also solving the problem that already exists today. And I'll talk a little bit about that.

I completely agree with what Rob Califf said, and that is this notion of, we are holding the bag as providers of care, and so when it comes to monitoring and these types of things, this is now our responsibility because it's nobody else's. And that's kind of this broken cycle that I talked about, and it doesn't matter to me if it's on the left side or the right side. It's still the broken cycle of trying to capture and record this information that's necessary to determine if these devices are working for our patients.

If you think about the ways to be able to monitor, it doesn't really matter before or after it all kind of looks the same if you just step back a couple of thousand feet. So, there's software testing when you've got a device, and as soon as you can get the bugs out, and you can throw some data into it, then you do performance testing. Typically, as was discussed yesterday, this was benchmarking and other different ways to solve this problem. But suffice it to say, you have to have some sort of answers or ground truth if you're going to test software that actually generates answers from data. So, you have to have the A, inside and then the outside to be able to test that. So, that's test data, and that's in my mind, standalone performance testing as it's defined by FDA today.

Then there's other mechanisms, and the one that I talked about yesterday that I'm not a big fan of is comparative effectiveness. But what I think is important is the concept of having subject-matter experts to be able to test that data to determine its accuracy. Whether you do that with LLMs on top of LLMs as models acting as humans, or you have actual humans doing it. The problem with doing this in a premarket way, you're doing it in an incredibly artificial way. So, the issue I have with comparative effectiveness and thinking of it as a drug, but it's not, it's a device barely it's really software, is this notion of now trying to AB test with and without devices and putting people in artificial situations to then test those devices and look at their performance. It typically just doesn't work very well and it's incredibly expensive and it doesn't get you much further than just the benchmarking before you deploy a product.

And so, if you think that something has got a higher risk and you impose comparative effectiveness, you're just slowing down the process of that product getting out, you're not really a bit of getting risk in my mind. As this product deploys to the site, now you've got deployment. So, the same product now is being dropped into a

facility, and that's where things start to change. So, all of this done before here, I know I'm reiterating this, is the notion of premarket testing. And now we have new data for the first time that's never been seen by this algorithm before, the product, and you have new subject-matter experts that have never been seen by the product before. And that's at multiple sites with multiple manufacturers. And that's the world that we exist in today times hundreds and hundreds of algorithms at hundreds and hundreds of facilities.

So, the challenge I see with premarket testing is what we're saying is we're going to do a standalone performance test. If you hit some bar of threshold, let's call it 80% for sensitivity specificity, then I would expect that this is going to work at all sites that it's deployed at. And the problem is it doesn't. So, if my site is functioning at an adequate number, do I really care that the algorithm is not working at other facilities? I really care about my facility. And if you do take a standalone performance test and convert that into a comparative effectiveness test, you really haven't moved the bar at these facilities where that solution is not working, and I'm still working, but I was working before with a solution and that was easier to get into the market.

So, what I would propose is that a standalone performance test is fine, but what's not fine is the inability to determine if that solution is going to work at these various facilities. So, on top of a performance test, this is what I was talking about yesterday, the notion of a clinical validation or a site validation. This happens today. Nobody deploys AI and it expects it to work. No matter whether it's FDA cleared or not, whether it has had a performance test, whether it's had a reader study without doing clinical validation because there's just no way to guarantee that it'll work. You'll find out sadly months later after you have a problem. And, you know, I've gotten pushback from the FDA, before explaining this, and they say, well, there's been no device recalls.

It's because this stuff is quite frankly not important enough to be able to do a device recall.

So, typically when it doesn't work for us, we just throw it to the side and treat it as if it's not going to give us the answers that we want and eventually gets eliminated. But there are major problems that are happening today because there isn't this concept of clinical validation. So, what does that mean? That means that you're basically removing that concept where you've got test subjects and test data in a pseudo artificial environment and putting a real environment now with real facility data, real subject-matter experts that have to make the determination of whether this is useful or not and do that clinical validation ideally before you deploy.

And because the product in theory will change through change controls and other mechanisms, and those happen today, that facility data continuously changes. And the one thing that I would try and stress I touched on briefly yesterday, these devices don't act directly on patients, oftentimes, they act on data coming out of devices that act on patients. So, for example, many of these are working off of image data coming from MR, CT, PET scanners, X-rays, etcetera, and those devices in our facilities are changing all the time. The output of that data is changing all the time. The protocols on those devices are changing, the resolution of those devices is changing. None of that gets tested again. So, we now are responsible for testing all of that data continuously and comparing that product to that data, and making sure that it comports to the accuracy that we require. So, that continuously happens today, and it's only going to get worse through GenAI because it's going to do more, and it's going to be stronger.

Also, the subject matter experts change their opinions and thoughts and different ways to determine if there's an aneurysm because the size changes by just metrics changing. So, there's all of this change that goes on, and that's this notion of the requirement for monitoring. It's really basically, in my mind, clinical validation that it either happens episodically or continuously, but it happens after deployment. And in my mind, the risk of the product should determine the content and the frequency of that monitoring that's required. And that is postmarket testing.

Now, this is this test harness that I talked about, this healthcare arena that us and others have looked at as a solution for all of these stages that we're required for if we're building product to deploy at our facilities or if we're buying and deploying product to be deployed at our facilities. And whether the manufacturer provides it or not doesn't really matter to us, we have to do it. And so, in this particular case, I'll go back to this draft reporting solution, and that is this notion of saying how do we monitor that component that is a medical device piece. And we do that by monitoring the difference between the draft and the final report, what it puts out, and what got corrected. And so, that's as simple as taking that solution that is the draft report creator, let the expert make the correction and create the final report and take the difference between the draft report and the final report and use that as the comparison, put that into the registry that was talked about yesterday at the Merritt College of Radiology, make that available to the site and make that available to the manufacturer, and also potentially use that to improve the product as well.

This is integrated into the ACR's Registry, which is a national solution, and I would argue that if there aren't a lot of monitoring solutions that exist today, but if the FDA said, these are the requirements of manufacturers to be able to do this, then I think

you'll see industries start to grow solutions because many people have created these solutions, but there just isn't a lot of uptakes. So, I think if there were a requirement for manufacturers to be able to do this, that I think you'll see that the industry at large, the public will start to create solutions, which is already happening, as they say.

So, the monitoring, the difference between the draft and the final report is what can make this product a solid and continuously solid solution, which maintains and improves its accuracy. And that is the basic, the solution. So, the notion in my mind, as I'll say, I just iterate for the final time, the notion of comparative effectiveness should give way to this much more solid solution, not just for general AI, but for all AI that we're using with the notion of clinical validation and monitoring. Thank you very much.

Dr. Bhatt: Thank you very much. I'll ask once more, Raaga Kanakam from Doctors for America. Are you here? Okay, we'll move on to Kerri Haresign, CTA.

Ms. Haresign: Great. Good morning and thank you to the FDA and the Panel for the opportunity for CTA to be here and present a little bit about the work we have underway on health AI and standard solutions. Just in case you don't know who we are, the obligatory corporate slide, CTA, the Consumer Technology Association. Our mission is to help innovators of all sizes grow. We have about 1300+ member companies and it is really important to highlight that 80% to 85% of them represent small to medium-sized businesses. Very important when we're looking at innovation, particularly around generative AI and health AI. We have a lot of internal resources including market research divisions and working groups, and where I spend my time in our ANSI-accredited Standards Development Department. We're also likely most well-

known for our trade show, CES, which is the most influential technology trade show in the world. We are really excited to be back in Las Vegas, January 7th through the 10th, we'll have a great digital health summit with a lot of highlights on AI and GenAI. Two specific Panels pointed out here. Again, looking at health AI startups as well as health AI in 2020. I think that's actually supposed to be 2032, so apologies for the typo, but we hope that you're able to join us there and learn more about these innovative solutions.

Inside of CTA, we have our CTA health division, which strives to increase the use of technology-enabled value-based healthcare to reduce healthcare costs and drive better health outcomes. We have a significant set of diverse stakeholders from consumer technology companies to MedTech to health plans and everybody in between. This group has been heavily focused in the last few years on all things AI, and continues to help our association's mission and scope in this direction.

And as I mentioned, we are an ANSI-accredited Standards Developer. I like to say that there's probably not a piece of consumer technology in your home that isn't touched in some way, shape, or form by a CTA standard. We're very well known for things like closed captioning and airplane mode. If you flew here, you probably used our standard in that process. But we have 31+ standards in the digital health space. And our AI standards specifically started in 2019. Because we are the Consumer Technology Association, we have focused on both AI horizontally, looking at AI across a variety of sectors as well as AI and healthcare standards. I won't read all of those here for you today, but I would like to highlight one specifically, which is our most recently published one, which is CTA-2125. This document looks at best practices for information disclosure and, as I've been listening to the various conversations over the last day and a half, I'm really looking at ways that we can communicate to end users

what's happening inside AI solutions is going to be very powerful from a standardization perspective.

We also have recently taken a little bit of a step back and said, how can we focus on our intentionality and standards development to help address many of the questions in industry and that the Panel here has been focusing on? And as a result, we recently formed a Health AI Planning Council, which is a multi-stakeholder group of health sector experts collaborating on a roadmap to identify where there's needs for specific standards in health AI. While we've been initially focused on predictive AI, we have started to start that parking lot list and figure out where there's going to be differentials and additional information needed for generative AI. But the goal of this council is to work with accredited standards developers like CTA, to identify and develop standards that are measurable, practical, adaptable, and consensus driven. We're really excited to get to work on this further in the beginning of next year.

While that's been focused on predictive AI, we've had a lot of different industry activities looking at generative AI, one of them being looking at education. So, we've developed a lot of education materials, such as our issue brief on reimbursement for health AI and innovative technologies. And we've been tracking GenAI, as I mentioned, to identify where there might be additional gaps in areas for standardization. And we've been working with other health technology associations through our health AI collaborative group to figure out what other education is needed and have done some great briefs, including what is a health AI document. A lot of that also continues to happen via events and bringing folks together. We have a handful already scheduled into 2025 on the AI-Powered Health Workforce, our Health AI Liability, and our third annual Health AI+ Conference.

So, at CTA, again, thank you for the time to address you all today, and we look forward to working with you on this important activity.

Dr. Bhatt: Thank you so much. One last call, Doctors for America, Raaga Kanakam. All right, with that, it concludes our morning speakers. I would like to thank all the Open Public Hearing Speakers for taking the time to address the Panel today.

As a quick summary today, we heard-- Today, actually we started with an echo of yesterday when Dr. Califf voiced a concern that AI outcomes are currently fiscal and that was echoed in our first speech. We heard a suggestion that consumer health devices with generative AI are somehow regulated to ensure patient safety from our video speaker. There was an emphasis on notifications to users of the device and procedures to remove those devices from the market when there are complications, and a strong statement to have thresholds for companies which repeatedly violate safety standards when it comes to generative AI. The AMA shared their additional emphasis on four areas: transparency, validation, integration into workflow and privacy and security, which we also covered yesterday. Specifically, they mentioned transparency mandates potentially via labeling, continuous testing and refinement of devices both in the pre- and postmarket settings to ensure safe and equitable deployment.

The AMA agreed with the Panel's belief from yesterday that AI for administration may have less stringent monitoring than those that include diagnosis and treatment. They also emphasize the need at this stage for human in the loop. I can hear my computer science colleagues, so I just want to say we recognize that human in the loop actually refers to the development stage of these technologies originally. We also recognize that we have taken that and now moved it to the post-development setting

where we're requiring humans to interact with the device. So, we know we've changed that definition. I also bring it up because it's a challenging definition to change because how do you regulate humans? And that has come up yesterday and again today. As we think about privacy, the AMA reminded us of the importance of opting out as a patient right.

Our colleagues from Deloitte addressed the multi-use challenges of generative AI. Unique challenges. They brought up that GenAI is more difficult to evaluate because of its qualitative nature. Errors can be hard to detect and difficult to control, and it's hard to control, to predict, and test all potential inputs, as those may change over time. They talked about the resource intensive nature, the compute power and human oversight. We have not here yet talked about compute power and its effect on the climate. I just want to note that we didn't talk about that yet. The Deloitte trustworthy AI framework is comprehensive mirrors frameworks that we saw yesterday from HCA and others. However, I'll remind the Panel that our job is using these conceptual frameworks and now offering concrete infrastructure suggestions to align with these local frameworks, and that's a challenge.

While we heard about the evaluation for misinterpretation and jailbreaking where the device expands its own reach, measuring biases, measuring human over-reliance involves a very different mechanism of postmarket surveillance. I challenge our team to think about those as perhaps two unique aspects to postmarketing surveillance. And then lastly, we talked about turning postmarket surveillance from a stick to a carrot, data drift mitigation, continuous verification, multilayer application, drift mitigation.

Dr. Dreyer then reminded us radiology has lived in the world of AI deployment without a monitoring mechanism and therefore the work we do here may help guide not only generative AI but also align with AI performance in real world in general.

And lastly, we are grateful to actually have the Consumer Technology Association here with us because as we've heard, many of these devices are going straight to consumers, and we do need to figure out how that fits into the work we're doing here today.

So, with that, let us do some questions from our Digital Health Advisory Committee. For the Panelists who spoke this morning, I'll ask you to keep your question brief. Please start with who you are addressing the question to and whoever we pull up first will then just orient questions towards that person and then move forward. Floor is open for questions. And please say your name for the sake of the transcriptionist.

Open Committee Discussion Q&A

Dr. Radman: Thomas Radman. Dr. Dreyer, we've been talking a lot about local validation, and as have you been in your two presentations, you had the slide today with the bar graphs, some green, some red.

Dr. Dreyer: Yeah.

Dr. Radman: So, in the context of FDA regulation, what do you propose would be the FDA's role if, for example, one site has inadequate performance while the others are performing okay, and what maybe would be some causes and appropriate response from the FDA or the sites?

Dr. Dreyer: I'm sorry, what was the question? I get the context, but what was the question, sorry?

Dr. Radman: What's the FDA's role in designing regulations? Is the purpose here?-

Dr. Dreyer: -Yeah-

Dr. Radman: -Should they recall a whole product if one site is--?

Dr. Dreyer: That is the question, right? I call it the Tazbaz Test because Troy challenged me with this a while ago. Is that what happens if some sites are good and some sites are bad? Do you recall or not? I know. If you look at it a little differently, if you look at when there was a recent cybersecurity meeting at the White House, it was like to figure out who is responsible to make sure that attacks don't happen. If it's the same thing with AI, who is responsible to make sure that these devices do what they're supposed to do? And if you look at the EU AI Act, there is this concept of a deployer.

Now, I know this is antithetical to what is it the FDA looks at today because your folks that you work with are manufacturers, not providers, and to step into provider space is questionably the practice of medicine. I just think times have changed. Like I said, these devices are not devices that are working on people, they're working on data, and so we have to keep the data pristine to make sure these devices are working independent of what the manufacturer makes.

So, I think it's time for the FDA to think about extending just a little bit beyond the scope of a manufacturer today - because that's what these devices do, and to people that are using these devices, and maybe we have requirements for things that we need to do. So, if site one needs to be able to monitor to prove that it's valid and site two through 99 aren't seeing that effect, is it not unreasonable if these devices are specific to

certain domains and actions and activities that they could be working at one facility and not working at others?

I would also throw in the concept that it's getting much easier for providers of care or anyone to manufacture these solutions given off the shelf high-end foundation models. So, you can imagine a number of facilities could be creating their own solutions that aren't intended to work anywhere else. So, I think again, the FDA would have to think about how to make that work at a particular facility and not every single solution.

Dr. Bhatt: Thank you. Dr. Rariy?

Dr. Rariy: Yes, thank you. Also, for Dr. Dreyer. You mentioned towards the end around the importance of creating this monitoring surveillance. Perhaps that looks at the difference between the example you gave was the draft and the final report. I'm curious to hear your thoughts around if we're thinking about generative AI specifically and once it's released out into the real world, it gathers more data, it improves, one would hope and perhaps think that it would be improving on the original model. And so, leveraging a monitoring that looks at the original draft compared to the final draft at that point in time may not be the best structure. I'm curious to hear your thoughts about what does that monitor-

Dr. Dreyer: -Sorry, excellent point, but I would just look at it. I try to separate the two. So, just to make sure that you would be monitoring both of them appropriately. The one is the output, independent of what the technology has done. So, I'm asking this device to be able to create a draft report, and it failed or succeeded based on a metric as to how much work that I have to do to correct it. So, I would argue that that still is a reasonable way to analyze that device independent of what's happened. I would add to

that fact though that what's new, that everything I've described is exactly what the American College of Radiology is doing to help radiologists to monitor these devices that are constantly seeing new data, but the devices are typically not changing unless there's process change control. What's changing I think here with generative AI, particularly with foundation models that are out almost in the consumer space that are being used for healthcare that are continuously changing is there needs to be another benchmark or a test harness to be able to evaluate those to make sure that they're performing at what it is that they're designed to do inside of that device. Right?

So, with what I'm describing, there are two different ways to solve this draft reporting solution. There are companies that are making a specific Large Language Model solution dedicated to the field of healthcare/radiology and have made these devices that are tight, fixed, controlled as a device. So, there are subcomponents of a device that gets used by devices. That's much easier to be able to use quality management systems and, you know, the things that we do today with devices.

I would argue though that there will be other solutions that take a different approach that say we're going to take a generic large language model that would do anything, and we'll focus that down to radiology, but we have to understand that that device was not intended for medical use, and it will continuously change. It's the obligation of that manufacturer in my mind to make sure that that component is tested continuously if they don't have control of that device.

But for me as the end user, I just want to test the end product to make sure that its inputs, outputs are functioning as they need to. And so, from a monitoring standpoint, the other thing I didn't add here is if the risk is high, I not only want to see

what the output came from the device and how it got changed, but also what the input information is too. I want all that recorded so I can kind of postmortem adjudicate anything that goes wrong.

Dr. Bhatt: Thank you. Dr. Soni.

Dr. Soni: Apruv Soni from UMass. I want to press on something that you mentioned both today and yesterday, which I totally agree with you, and I'm on board that clinical validation is important, needs to happen locally, not just pre-implementation but ongoing clinical validation. But one of the things that worries me is this de-prioritization of comparative effectiveness.

Dr. Dreyer: Yeah, yeah.

Dr. Soni: And I think there may be an opportunity to do both instead of deprioritizing comparative effectiveness, and we may need to think creatively about how we are doing that comparative effectiveness. Totally agree, randomized control trial is not the only answer, especially because by the time you do a randomized control trial, the solution may have changed.

Dr. Dreyer: Yeah.

Dr. Soni: But there are synthetic control trials using synthetic approaches different than synthetic data and would love to hear the distinction about that. But that can allow for ongoing continuous postmarket monitoring of comparative effectiveness in addition to clinical validation?

Dr. Dreyer: Yeah, let me be very specific. I'm so glad that you raised it the way that you did, because I do think that it's important to do something beyond the scope of just

a standalone performance test as you're describing, but it's almost as if we stepped into the drug world and said what did they do when things got complicated? Let's drop that into the device world, then extend that another layer to say we're going to use it for software.

And now let me just describe what it is that's not working is that if I take highly paid physicians, radiologists, and ask them to perform a specific task, which is not what they do every day, I'm asking them to look for a stroke, they look for everything all the time, and now I put them in a white room with a cup of coffee and I have them read a hundred cases, and then I do it with AI and without AI and that somehow is a real-world environment. It is nothing close to a real-world environment, but it costs millions of dollars, and it wasted years of innovation. That's what I'm describing as a complete waste of time compared to a standalone performance test.

Dr. Soni: Just as a follow-up, I think there is potential for natural experiments that are happening because the implementation of these technologies varies so much and my challenge for all of us is to think about how much can we leverage those natural experiments and real-world evidence strategies to generate the effectiveness data because validation and accuracy testing helps with the safety aspect of it, but does it make a difference? It is still a question that we need to be asking.

Dr. Dreyer: I think there is something to be determined that could be like if you took a look at a site validation, the problem with what I've described is it has to happen at every site. I don't think there's a way to get away with that happening at every site from what we've seen with the AI today, but I will add that if there's something that could be done a priority or premarket that solves 80% of those sites, that would be wonderful. I

just don't think that we have that tool in our arsenal today, but I like the way you're describing, there are things that we could probably create. Yeah.

Dr. Bhatt: Thank you. Dr. Botsis?

Dr. Botsis: So, what is the actual cost of the monitoring process you described? Do you have any estimate on that as well as the actual cost of going back and probably retraining the model, changing the technology, especially in the case that the clinical validation fails despite your expectation for that not to happen?

Dr. Dreyer: So, I'll give you some numbers just to look at it. I'm curious, Nina, who has a larger practice is responsible for and is talking later can probably give you more insight, but just as an example, we have about 15 hospitals and when we look at a product solution to determine whether we can onboard it, we do clinical validation, we have to do clinical validation, and we estimate internally that cost is about $50,000 to do. I would argue though that every one of those we do today is a one-off and that's why this whole arena concept was to say we can't keep doing this at $50,000 a click. Let's build a solution that we can now do a much cheaper way to evaluate these. So, if there is a way to do it episodically, it depends on how much data you're pulling in, how frequently you're doing it.

So, if you're just doing it episodically, you could dramatically reduce the price, but I would argue that the deployment of these devices that cost quarter of a million to a million dollars to deploy inside your facility, if you don't test them for that amount, then you're going to be wasting tremendous amounts of money and harm patients in the process. So, it's a necessity, even though there's a cost. The question is, and I think you

raised this earlier, who pays that? Whose burden is this? But in light of nobody having this, it's our responsibility, and we have to do it.

Dr. Bhatt: Thank you. Dr. Elkin?

Dr. Elkin: So, I'd like to push back on this a little bit. I think when you said that you could just compare the judgment of one clinician against what the AI said and that's going to be the gold standard for whether it's right or wrong, I think it's a very poor gold standard. I honestly think that the evolution of thought that goes into medicine happens over time. You don't know if the-- I'm sure it's a Venn diagram, but is it 90% of those differences the AI was right or is it 10% of those differences the AI is right?

I believe that as you look at clinical care, there's a natural evolution of thought over time that happens. You do testing workup strategies, you further evaluate the patient, and you come to closure at the end of an episode of care, whatever that episode is and how we define it. And it's probably at the end of that episode where hindsight's 20/20, and you can go back, look at things like those differences, but also look at the final diagnoses, the discharge diagnoses, the outcome of the patient and then the quality of the care that was given by some other measures, NSQIP or others that have been validated.

And at the end of the day, we could at that point have a pretty good lens. If we look at the bio-surveillance literature, which is quite robust, this is how people would recommend that you go about looking at this kind of an evaluation. The kind of point estimate that looking at the variability of doctor's judgments a good one cardiologist reminded me, if you look at echo ejection fraction estimates by cardiologists blinded to each other, it's pretty wide confidence interval, right?

So, I mean when you think about everything that could be in a chest X-ray or a CT or an MRI and whether the same exact text would come out of a report, the differences are going to come out whether they're meaningful or not meaningful, whether the physician is correct or the AI is correct and all these things will be sort of hanging chads if you will.

And so, I think we need to think this through a little more carefully before we settle in on a bio-surveillance strategy that we think is something we should recommend to the FDA.

Dr. Dreyer: I would just add-- Can I comment? Well, first, I would say the concept of we're just having one physician look at one output is not the case. We do 15,000 exams a day, AI runs on all of those and we have 500 physicians. We're recording all of that information. So, over time we're doing exactly what you're saying is this kind of single signal has got a lot of signals that come out.

We look at that over a long period of time to determine if there's decay inside of the accuracy of humans to humans, humans to devices, devices to devices. And what we see is that AI fails, particularly when we change a protocol on a CT scanner, not because a radiologist uses a verb instead of a noun and we can identify that signal and that is when things need to be pulled out of the system.

To your point about outcomes, and actually what is ground truth and what you should be measuring, radiology has always suffered from trying to decide: does this CT save the patient two years from now? It's very difficult to determine that inside of a full care process, but what we can do is measure inside of our activities in radiology whether the output of our product, which is the report, is at the accuracy that it needs to be.

Dr. Elkin: So, years ago I will send you this, I published a paper on who makes the diagnosis. It was so much fun. We had the radiologist told us they made all the diagnoses, the laboratorians told us they did, the clinicians told us they did, you know, everybody came up with their thing. So, we actually took 248 cases, came to the hospital at Mayo in a row, and we had a strict protocol on how to do it. It turned out, just to give you the punchline, about 50% of the diagnoses were the clinician, about a third were radiology, 9% for pathology, and then it was a smattering of all the other modalities that we had.

But the truth of the matter is that there is a process involved, and we need to understand that process. I still don't think a point estimate by a single person for a judgment. So, if you had 10 radiologists that looked at the same scan, and you coalesced what they said and that was the gold standard on what the AOB compared to, I would think that a better protocol, but you can't tell me that there isn't variability in the quality of the clinicians that work with you because they're all bizarre, right?

Dr. Dreyer: So, we do peer-review between physicians. The point is that we now do peer-review between physicians and devices. The other thing I'll add though just to close is that these devices are outside of our control. So, they're used by clinicians, somebody trying to decide if there's a pulmonary embolism that they're not sub-specialized in, and yet we're responsible for that action as well because the device was scanned and the patient was done in our area. If we don't keep track of all that's being done even outside of our area by clinicians that think that they know how to interpret the AI, then who's going to do that?

Dr. Bhatt: Thank you both so much. It's a great discussion. It's probably one we could have for half the day. I'm going to put in my two pence because I can't help it, but I think the clinically significant differences are a reasonable place for us to start when we start thinking about how one regulates and creates a framework because the other differences you pointed out cardiologists, and you're right. Boy, we differ in a lot of different things. All right, we're going to go to Diana Miller and then Steven Posnack next.

Ms. Miller: Yeah, my comment is related to what you guys said because-- Thank you for clarifying that your monitoring strategy involves monitoring both of the inputs and of the outputs. So, my question was what do we do when we realize the monitoring inputs on a clinic show that the input from that clinic data knowledge is not good? Then you go to that regulating practice of medicine. So, what do you do if the inputs are wrong from onsite? Then what's your proposal to do?

Dr. Dreyer: So, historically, let me speak historically because this is real world. What happens is we find inconsistencies, and you can't just say that a device that's running with 8 million patients and 2000 devices with 500 clinicians went from 92% to 91%. You rip into that data and try and figure out is there anywhere when it went from 95% down to 5%? And the answer is typically you find that as a yes, and it's the input data that has failed because these devices don't change. When these GenAI devices change, that's going to be a new problem.

But today what happened is someone just changed the slice thickness on a CT scanner and these devices weren't trained to be able to support that. So, then the question is you either take that device out of commission for those scans or you change

the protocol of that device and that's what typically happens on a regular basis. But unless you monitor, you have no way to know that and that's how you find out that the device says that there's a fracture when there's not, because the humans can deal with the change in the slice thickness and these algorithms cannot.

Dr. Bhatt: Thank you. Steven?

Mr. Posnack: All right, long-time listener. First-time caller. I feel for our FDA colleagues, you know, policy changes are all about trade-offs and choices, and so we've heard a lot of nice-to-have criteria I guess I would say at this point, but not necessarily as many must-haves and which ones are going to be the most impactful relative to their burden. So, this is for everybody, I know you're on the hot seat, but this may be relevant to you too. I think we've seen-

Dr. Dreyer: -I like that-

Mr. Posnack: -dichotomy between pre and postmarket-

Dr. Dreyer: -Yeah.-

Mr. Posnack: -and there could in theory-- And we've seen some shift from everybody's comments to the latter trying to move things in the postmarket space.

Dr. Dreyer: Yeah.

Mr. Posnack: There could in theory, especially in the software context, be a step in between where there's some intermediate testing, and I'd almost posit here that with software there's premarket and then there's prescale, because some of the issues that we've been discussing are matters of scale, and so I don't know implantable devices all that well, but once they're approved, nationwide availability.

Dr. Dreyer: Right.

Mr. Posnack: And that's the scale. And so, with software here, we do have an opportunity potentially to look at it from a prescale or some type of intermediate testing. All of us are old enough to remember that Gmail was in beta for five years. And so, you know, there could be some to your presentation as well as others where there's some type of in-the-field assessment for some period of time-

Dr. Dreyer: -Right.-

Mr. Posnack: -that isn't at a postmarket scale nationwide availability, and was curious if any of the other presenters as well have some thoughts on that.

Dr. Dreyer: I mean that's what we do. When I say clinical validation, that runs for 90 days, six months before we do it. So, it's exactly what you're describing because it's a test harness that's controlled, it's not releasing, we're not affecting patient data. We have a managed way to be able to look at this, but do it at scale.

Mr. Posnack: That's after the FDA process, though.

Dr. Dreyer: That's absolutely after it's been approved.

Mr. Posnack: I'm wondering if you think that there's somewhere-

Dr. Dreyer: -Before-

Mr. Posnack: -to marry.

Dr. Dreyer: I mean, that's also-- We have a CRO as well, and so we do premarket testing. It would be interesting to kind of devise a way, and this is my point about comparative effectiveness is there's something else that could exist there to your point that's before the market's released that is tested at scale. We've all talked about

collecting data into a central area and testing all of that, and that is this notion of arena. If you could get people to contribute data to a place, I know the VA is talking about doing the same thing, to be able to test this out to inform the FDA of the results of a scale evaluation of clinical validation, not a, you know, dark corner comparative effectiveness test. I think that is kind of the solution that would solve the 80/20 rule of a lot of places.

Mr. Posnack: You want like four out of five Tazbaz's tests to agree.

Dr. Dreyer: There you go. That's it. I'll just, the other thing I'll add because I left this out, it's not obvious is if this reader, this draft reporting thing, if this was done by a non-specialist, if this was just read by an intern, they're not experts, so their data of before and after is not going to be valuable. So, this predicates the need that, I use this analogy, I appreciate the fact that there's experience in the automotive industry that you had mentioned, is that I think if there was an autonomous bus solution that someone could drive a bus autonomously, I still think we would probably say you have to have a license to drive that bus. You can't just have a driver's license and hop in a bus. It's the same thing that I see in radiology today is that these devices should be used by people that are competent at reading the data regardless of the value of the AI. And if you require that then you'll see much different risk dropping and the use of the device and much easier ways to monitor it because that's their job. No radiologist is going to sign off on a report that they don't agree with because they're liable. I don't think that exists in all areas of medicine. It's the same for cardiology, right? If the cardiologist is responsible for this, and you add AI, they're still responsible for it.

Dr. Bhatt: I love this conversation. We're going to go to Dr. Rariy and Dr. Kukafka next. I'll just bring up that last point you said challenges the notion we started with yesterday morning that the advent of generative AI is going to increase access to care in the areas that need it the most that are suffering from a lack of adequate providers because of our shortage. And so, I think as we, the Panel, think through this, we have to think about when experts are using versus when non-experts are using and perhaps that is just a point for us to discuss post-lunch when our discussion comes up.

Dr. Dreyer: I think there's two ways to think about that though. I would say there's an incredible shortage of radiologists today and everyone is hoping that AI will kind of ease the load for us so we can be more efficient. If there was a way to improve our efficiency through AI and I think GenAI has a way to be able to do that, then you're going to see availability, much broader availability of even specialists to do more work and to do work beyond the scope of what they can do today. The other approach is to take AI and make it more autonomous and deliver it to those that have a lower experience in a particular domain. That's a much different risk though that we're talking about.

Dr. Bhatt: It is, and it requires a different infrastructure. All right, Dr. Rariy and then Dr. Kukafka.

Dr. Rariy: Thank you so much. I think that's an excellent segue to the point I wanted to bring up to you, Dr. Dreyer and other Panelists as well, but just pointing out that obviously MGH is a huge well-funded conglomeration, and so we need to consider that and to consider what the capabilities of smaller facilities, smaller practices like FQHCs or community specialists would be and what requirements the FDA would put

on those companies as well and balance that cost-effective, if you will, safe approach. I would worry and do want to bring up that obviously if you're having the large facilities being able to monetarily develop these products, including the infrastructure that is required, we absolutely do run the risk of worsening health disparities of not being able to capture the real-world data on particular populations. And so, when these generative AI devices are deployed in specific populations, we certainly need to be cognizant that creating a mechanism to ensure that many individuals can participate in this innovation and development is going to be important.

Dr. Dreyer: I would add to that, and this was the point that might've gotten lost just in the speed of presentation around the American College of Radiology's platform for assess AI, which monitors AI. They've also, and other JACHO and others monitor devices, medical devices, radiation, ionizing radiation devices for their safety. And so, accreditation is what the American College does. That's why they're connected to 40,000 facilities, and it's a charge per device of a few thousand dollars to just maintain that. So, I do see if AI could happen in the private sector and just it's required by government, but then the private sector can come in with solutions. And that's what I see solve the problem for that is CMS won't pay for a CT or an MRI unless it's accredited, and the accreditation happens by JACHO, the Inner Society, American College of Radiology for a few thousand dollars to all facilities. I think it's the same thing with AI that's going to have to take place.

Dr. Bhatt: Thank you so much. Dr. Kukafka, who is your question for and your question?

Dr. Kukafka: Hi. We stay with Dr. Dreyer since you've been on a roll.

Dr. Dreyer: Sure, yeah. Let's go.

Dr. Kukafka: So, as I'm thinking as someone who does a lot of evaluation of digital health, and we talked about comparative effectiveness and problems with that, which I agree in this context, and there was some mention of RCTs through and there were good reasons not to do RCTs, but I'm wondering in terms of collecting rigorous data for real-world evidence, do you think there's a role for pragmatic trials?

Dr. Dreyer: Such as?

Dr. Kukafka: So, pragmatic trials are different than, like RCTs because they are designed to adapt to the real world. So, it's not like you hold things constant-

Dr. Dreyer: -Got it.-

Dr. Kukafka: -like you did, you know, with the RCTs, you keep them-

Dr. Dreyer: -blocked, fixed-

Dr. Kukafka: -fixed and you can't make changes and there's a lot of things that get deployed or implemented into settings, and we realized that change happens-

Dr. Dreyer: -Yeah.-

Dr. Kukafka: -and there's a lot of variability, especially like a multi-site pragmatic trial where you would actually, you would be doing the same thing that I think you're describing, but I think framing it as a pragmatic trial and establishing those guidelines of pragmatic trials will help us provide more rigorous evaluation. So, I was just interested in your thinking about that. Is it feasible? Is it--

Dr. Dreyer: Yeah, it's back to your point, too. I think that there is something that's not comparative effectiveness-

Dr. Kukafka: -Right.-

Dr. Dreyer: -that can and should be done. I would just add to the fact that I can't tell you if it's accuracy or style. I mean, to your point of why this won't be acceptable, but I bet if you do do that, and it's an 80/20 rule, and it does the Tazbaz 80% of the time rule, right? The 20% are not going to be happy, probably not because it doesn't work, but just because it doesn't work for them because of their style. To your point about the difference between interpretation, some people just won't find the solution acceptable, but it doesn't mean that it's not accurate, or it's not safe. So, I think that I would be a strong advocate of saying, you get as much as you can pretest probability to get accuracy out there, but don't count on that as solving every problem for a deployment. It still may not work at facilities, and that's okay.

Dr. Kukafka: It would be good, though-- That's why I'm looking for something with more rigor because-

Dr. Dreyer: -Yeah, yeah.-

Dr. Kukafka: -20% is not okay if it's high risk, right? So, knowing under which conditions it doesn't work would be helpful and that's what you would get for a more formalized approach than saying we accept the 80/20 because we don't-- 20 could kill a lot of people.

Dr. Dreyer: No, I'm sorry. When I said 80/20, what I mean is it won't work at 20% of the facilities, and they won't deploy it.

Dr. Kukafka: Well--

Dr. Dreyer: But I mean, if you're shooting for 100% with something that hallucinates, it's going to be incredibly difficult.

Dr. Kukafka: I don't think it's shooting for 100%. It's identifying the conditions in which we know it won't work.

Dr. Dreyer: Yeah. But if it's any measure of what happens today, if a device is shooting around 99% accuracy, it's absolutely incredible, and it's still making an error every day. And that's kind of the way AI is today, at least the way it is. And I don't see generative AI being more accurate. I just see it more broadly useful.

Dr. Bhatt: I love this conversation because it echoes a little bit to premarket what we were talking about yesterday with having the devices actually pre-state. Hey, these are some populations we tried, situations we tried where things did not go as it does for the thing we're asking the indication for. And I wonder if in our discussion later we extend that concept and talk about it in the postmarket setting where there is some form of reporting that we are working on this that may be challenging from an industry perspective, but maybe something to think about and if such a state that Steven was suggesting as a prescale state exists. That might be a great place to say this works slightly better in the following conditions at this time. So, I think that would be interesting. Thank you. Dr. Kukafka and Dr. Dreyer. Dr. Shah, your question and who is it for?

Dr. Shah: Yeah. Hi, Pratik Shah. You can take a seat. Take a break. My question is actually-- I want to invite other three presenters like from Deloitte and Dr. Shrestha to the podium, and I'd also want to make some comments about and just have a quick discussion and a question for you. So, I think the site validation question, I think we

should really think through this very importantly and hear more perspectives, especially from the FDA because eventually if you have a site-based validation, we are kind of committing to a post-hoc situation where everything is done post-hoc after the model has been deployed. So, that's one thing I want to point out there. And in that case, we did have a pre-cert program at Digital Health Center of Excellence. And can somebody comment that do we still have the pre-cert program and how relevant it would be for getting sites approved or in line with what we are looking for? That must my first question, anybody can answer it and Troy, you can--

Mr. Tazbaz: Yeah, so also for the record, the Tazbaz 80% rule is not an FDA official position, so I just wanted to get that on the record. We currently do not have a pre-cert program and pre-cert was not necessarily used officially for postmarket. It was mostly around the premarket side.

Dr. Shah: Okay, so that's great to know. And the thing that I want to also point out, which I was hearing, is that we are-- What we are talking about the communication around it, we're talking about a collective intelligence problem versus a narrow intelligence problem for validating these algorithms. So, a collective intelligence problem is that we have multiple clinicians validating a set of algorithms and their collective intelligence is going into clinical decision-making and validating the algorithm.

If we switch to a site-specific thing, then the collective intelligence distribution of clinicians becomes narrower and limited to that specific site because that's a very important thing to point out that eventually we are looking at a situation where we are tapping to Peter's point, that we are tapping into a situation where very narrow sets of

clinical decision-making is going to happen at every site, and we will not know what the broader clinical community would think about it. So, the collective intelligence value of site-specific validation is going to be reduced.

So, in that framework, my suggestion would be to do both, have a site-specific program like pre-cert something where people can do it, and then have a broader collective intelligence program where multiple sites can also put input into the same algorithm and can help the clinical decision-making move forward so that we don't lose on the collective intelligence of the medical community for the decision-making.

And the third is that clinicians, and I'm also a Professor of Computer Science and Electrical Engineering, so clinicians often operate in a POMDP format, Partially Observable Markov Decision Process, where they really don't know a lot of decisions they're going to make, but they have a notion of what's going to come up in the future, right?

So, I think that no AI is going to be able to capture because a lot of that is from intuition and clinical signatures and I think that's why I think we need to have a more broader community if we go with the site specific validation model to put input into the things, and I really like Dr. Kukafka's pragmatic trial idea and that goes to doctors per this initial point about real-world validation because a lot of the platform adaptive trials, you can randomize sites into different arms and have different arms running clinical trials together without running a massive site. Obviously, for that you have to prevent confounding because you cannot randomize sites to do two different arms if they have different devices, different scanners, different clinicians, but you can still study the impact of this site-specific thing we are talking about, which is very critical. If we

decide to go with the site-specific model statistically there are going to be a lot of questions about who goes to which site, how do we randomize patients, do we randomize devices? How do we statistically explain this model, right? I really would encourage to consult a good statistician if you go down the site-specific road to prevent confounding.

Dr. Bhatt: Thank you so much. As you're answering that, Dr. Shrestha, I see you walking up to the mic. I just want to add a little question to get concrete while we have you, which is you talked about continuous validation and if we're taking a single device that is generative AI-enabled and putting it in multiple sites, that is something that could come with the device if you will. That would help us because we're not going to have the right data scientists, clinical informatists at all of the institutions the way Dr. Rariy was mentioning. So, could you say something a little concrete about how that process could work as you're answering this question of continuous validation at the sites?

Dr. Shrestha: Sure thing. Thank you. So, I'll try to address these and please definitely in advance, sorry if I get anything mixed up. The first point on the pre-cert program, I believe the current tool that we have at FDA is the excellent PCCP program and kind of the way that I envision that from a software engineer mindset, it's like test-driven development, but you're getting your tests pre-approved for FDA in your premarket submission. I'm always going back to the premarket submission because I think that's the chance, that's the moment where we can convince, in industry, to do the things there.

Otherwise, the market will find the efficient pathway to market and usually postmarket analysis after that first FDA clearance is likely going to fall through the

cracks there. So, I think something like a PCCP where even that can be phrased as a carrot rather than a stick, we say that, okay, this is something that you can go to FDA get a clearance sooner perhaps, and this also ties in with the partial kind of soft deployment mechanism there is perhaps your current algorithm isn't tested enough for nationwide deployment, but with a PCCP you can get clearance now you can start generating revenue now, and you can get that next round of funding which everyone cares about. But you have a pre-certified test-driven development type plan where you can say, once we meet these criteria, then we can roll out to the rest of the nation and this is key again without having to go back to FDA.

So, I think that's how we can leverage the PCCP to solve some of these issues that I've been hearing. As for the narrow clinical side problem, I think that could be something like a stochastic sampling where it doesn't necessarily like the burden of proof for that doesn't necessarily need to be put on the customer, but it could be transferred onto the manufacturer and again, in a way where in the premarket, perhaps this means in the standalone performance test, we need less samples because we have assurance that hey, we're going to get some validation in the postmarket.

So, that could be a way to kind of pitch that, so to get more compliance on that. So, like a stochastic sampling where images are fully de-identified, sent back to the manufacturer, and then in that case we can have a bit more control over what's there that can even be phrased is like a verification test that goes ongoing, which is I think another way in the premarket where FDA can have some insurance that this verification test is going to be run in the postmarket and that tends to capture, I think what this is. If it is a verification test, it's running the postmarket, and we're using the same terminology that everyone's used to already.

And as for the fully automated validation how that could potentially solve the healthcare disparities, I don't think the fully automated data synthesis and validation is really like the technology is not quite there yet to do that. This is a little bit kind of forward-thinking, but I think eventually it will be, and that's a case where I think you could have that if properly validated, if those models could be validated for that specific narrow use case, then that could significantly decrease the cost for these types of studies. I'm a terrible salesman, but I can sell for free and if I say, hey, this thing can be done very cheaply and for free, that's something I could sell.

Dr. Shah: That's a quick clarifying comment on that. I think for narrow use cases like radiology reports which are generated by large language models, it's totally fine because we are underwriting a very low risk as the provider, but if you're taking a clinical decision-making which recognizes tumors or something more significant, that's going to be a much more-- Because the radiology example that we've been hearing all morning is a very defined problem for a very defined LLM generating some reports, and that's fine. Of course, it has nuances and complexities, but there might be other bigger issues if we validate a site-specific algorithm or site-specific paradigm that people might use the site-specific framework to put things that are more complex than the radiology use case we have used. I just want to point that out. The radiology use case is great, but there might be more complexities if we go down that road.

Dr. Bhatt: I like that you bring that up because I think we do have to recognize, we also talked yesterday about multimodal inputs, and that's something that's very different than having a select image and a select definition. So, I like that not only for your comment about tumors, etcetera, but just thinking about multimodal in general. Dr. Shrestha, our Committee has one more question for you. Could you clarify the

distinction between generating synthetic data and the pragmatic comparative effectiveness approach of synthetic controlled trials?

Dr. Shrestha: I'm actually not familiar with the pragmatic synthetic controlled trials.

Dr. Bhatt: Okay. Dr. Rariy or Dr. Soni, I don't know who passed this, sorry.

Dr. Soni: Sorry. Just because I think there is often that conflation that happens between the synthetic data that's been created and then the approach where you're synthetically deriving a group of control arm. And so, I was curious whether or not you can differentiate between the proposal that you had of creating that synthetic data to help generate some of the evidence for the accuracy and safety versus the comparative effectiveness method of comparing the outcomes.

Dr. Shrestha: Sure, sure. I think the way that I use synthetic data is more akin to data augmentation. I'm not sure if that helps clarify where my vision is that we can use the foundation models that have been trained on a much kind of a global scale, right? And as those foundation models advance those presumably would be trained on the most up-to-date data.

So, data drift can be captured in the weights of those foundation models on a global level. And then if we use that foundation model to then data, augment data that is collected on site on a nightly basis, weekly whatever, but some kind of frequent cadence, then we can capture the global data drift along with the local data drift all in one setting and again, in a fully automated manner, which I think there'll be a lot more postmarket compliance to do something like that versus something a bit more burdensome where you'd have to send the data back and get it manually annotated and stuff that is less likely to be done.

Dr. Soni: Could you imagine there are specific sites that have the unique capability of doing this, or does that sit with the manufacturer themselves?

Dr. Shrestha: I was thinking that this would be the manufacturer. Pretty much my whole presentation, I was thinking that that could be something that the manufacturer would do as part of their ongoing verification of the product. And it's really not that much different to surveys that you have on your app. You know, how are you doing? And its common practice for websites to send back telemetry data about performance metrics and other things like that. So, you know, I think it just makes business sense too.

Dr. Bhatt: Great. Dr. Khubchandani, you had a quick question, and then Dr. Radman, Maddox, and then we'll close.

Dr. Khubchandani: Yeah, I just have a broader question about maybe for the industry representatives, many of these devices, predominantly cardiology, radiology, neurology, and we don't see many devices, AI, non-AI for urology, gynecology in an aging nation. I wonder if there's any focus on prostate health or such AI devices, and why this affinity just for radiology and cardiology? And no offense to Dr. Bhatt, but--

Dr. Bhatta: I would defer to more of the clinician in the-- Here, not industry-related expertise there necessarily, but--

Dr. Shrestha: Sorry, what was the question? Why is cardiology and also radiology significantly over-

Dr. Khubchandani: -overrepresented in these devices and with the aging nation few innovations in urology or gynecology, which is an equity issue as well.

Dr. Shrestha: Gotcha. I think one thing that really comes to mind is just the data availability. With cardiology and radiology, there's a lot of imaging data that is available, which I think historically has been kind of the main focus of AI, of non-generative AI. You could also argue that wise dentistry is not in there, I think because the standardization of DICOM and things that haven't happened yet inside of dentistry. So, that's another barrier to entry there. Yeah, that's the main thing I can think of, top of mind.

Dr. Bhatt: It's a great point. Yep, Troy? Absolutely.

Mr. Tazbaz: Yeah, so we do studies of that, and it's a great question, but it's always around supply-demand issues, right? And to Dr. Shrestha's point is that data availability becomes a big driver around where people invest. And so, when you think about data in the context of hospital systems, 50% of the data is pathology data, but yet there were no standards like DICOM we've seen in radiology. So, a lot of investments didn't go there. Radiology then represents probably somewhere around 25 to 30% of this is the data size. And then you had a smattering of other things like Omics data and EHR data is a very tiny little sample of that. So, because the data had been standardized through DICOM standards, then it was a little bit easier to apply AI to it, right? From a training perspective and then of course testing and validating.

Dr. Bhatt: Thank you so much.

Dr. Dreyer: Can I just add a point about why radiology?

Dr. Bhatt: Yes. And then we will go to Dr. Elkin and Dr. Maddox. Oh, sorry, Dr. Radman first.

Dr. Dreyer: I just go back to the slide that I used yesterday where about 80% are coming in as CAD-T, and I'll just emphasize the fact that that's triage. It's not a really functional AI tool, but it's a low bar to get through FDA that manufacturers have used. So, uptake is just not taking place. But I would say that, to Troy's point, you have to have data to be able to train as well as use these algorithms. And medical imaging data is pretty robust and consistent. That said, it doesn't have to work in radiology, it should work in other disciplines like urology, gastroenterology. It's just that the device process of getting approved through the regulatory process today is not amenable to that. And I don't see that changing with GenAI if we keep the same process.

So, I say risk is good. The process that's used to determine risk is good, controls are good, but the existing ones are not good, and they're not going to create devices that are going to create this equity, resolve equity problems, and give access to other specialties.

Dr. Bhatt: Thank you. Dr. Radman.

Dr. Radman: Okay. I welcome Dr. Dreyer to stay up there. This may be applicable to him, but I will welcome anyone else to answer. Really, two questions: one, and I think we've started discussing this, but my understanding of using, you know, we use the term real-world data, we use the term naturalistic studies, observational data, those aren't as robust compared to the gold standard double blinded clinical trial for a host of reasons, Electronic Health Record data is really non-standardized, and there's a lot of statistical challenges, I guess we'll say like multiple comparison errors. You know, you're talking about doing tests nightly, for example, that has a whole host of statistical challenges, and putting that in the framework of what the FDA has to work with, they

have certain really congressional mandates that are going to be very difficult to change like this whole paradigm of substantial equivalence, which is for the Class II products as you mentioned yesterday, is like the predominance, what we're talking about here with GenAI.

Dr. Dreyer: Yeah.

Dr. Radman: So, how do you use that kind of, let's just say call it real-world data to draw comparisons to a device that's already on the market that would be the predicate and a substantial equivalence comparison? And this could include devices that are, you know, pre- this discussion that only were evaluated in a premarket setting, and you have a 510(k) summary that shows some performance in maybe more of a double-blinded setting. And now we want to-- How do we compare good or bad performance in the wild setting? If you will.

Dr. Dreyer: Let me just start by-- I've spent countless hours trying to figure that out specifically for the existing devices, right? Knowing GenAI is coming, but even just that there's going to be many, many more of these devices next year than GenAI is coming out next year and this problem is not solved for that. I would just say, I would challenge this group to ask the question, are these things really devices? Like you said earlier, what Bakul said, I don't think these things are devices. I don't think that you're going to work through a paradigm of a round hole getting this square peg into it. It's just not going to work. And you're going to keep making things that don't work, and it's not going to create products that people are going to use, and you're not going to help patients, which is the whole goal.

I would just-- I don't care if you have to go back to Congress, I know this isn't simple. I'm not saying it. Someone should be thinking about we probably made a mistake when we turned software into a device and now software is incredibly intelligent, and we're still calling it a surgical knife.

So, I just don't think that is the right paradigm to answer the question. But the best I can say is that now you take, as was mentioned earlier, this is easy for medical imaging because a CT of the abdomen is pretty much always a CT of the abdomen. Clinical notes are never the same. So, I don't care what you tested on at hospital A, B through C are not going to have the same kind of notes and the device is not going to work the same way. So, I don't know what you do, but I don't think you treat it like a device.

Dr. Bhatt: Thank you. Dr. Elkin and then Dr. Maddox, and then we'll close this section.

Dr. Elkin: So, one of the concepts I think would be helpful to this discussion that we really haven't brought up is that there are different levels of risk with different errors. We have high-risk errors, we have medium, and we have low-risk errors. You know, this could be in radiology where you've missed a tumor or a PE, even worse, or it can be as benign as someone didn't note that there was an old-healed fracture in an ancillary bone. And so, I think there's actually been groups that have worked on this problem so that the information does exist on how to classify some of these things already. I would make-- And so I think we ought to think about that as we define our bio-surveillance program so that we can understand the level of risk and the level of harm by any particular set of errors.

So, just as a comment, because we do a lot of work with EHR data and there are observational data models and there are standard ontologies that can be used in order to standardize the information that's coming out of even clinical notes. And those can now, they've been used over billions and billions of patients worldwide already. So, this is not just me saying this, this is work that's been highly validated. The NIH uses it and has a policy on it already. And so, there's opportunities for us to use those technologies, which are standard in the country, to actually do this kind of bio-surveillance work.

And I think that perhaps as a group, somehow, we should have some pilots looking at that to try to tease out what the level of accuracy of those instruments are for the purposes that we are defining in our postmarket surveillance strategy. I think that could be a very nice validation for whatever comes out of the Committee long term. Thank you.

Dr. Maddox: I'm just going to make a comment, so feel free to have a seat. Appreciate very much your all's expertise. I think something that we want to take into our conversation later on today is we need to stay cognizant of the real world here. We are all biased, I would say, towards regulation and research, and data. And I have in my mind all the rural hospitals in our country, and they are like, "Look, you guys can talk all you want in the ivory tower. I don't have the data, the people power, the ability to submit any of this. Yet I am treating, I'm up to my neck and treating people who need my help." And in my world of cardiology, the AI we've been using for decades is that printout you get from your EKG. And I'll tell you right now, automation bias will occur. They will rely on these tools autonomously. You're going to have a very busy ED or primary care doc in a rural location who in the EKG case needs to know if it's an ST-elevation MI. That's critically important. And if it's not accurate, they're not going

to sit there and validate it. They're going to accept it and then move forward from the clinical decision point of view.

So, I'm not-- I just think we need to think a little bit about boundaries here. So, where do we think the boundary, knowing it's going to have to be a trade-off, is between regulation, data collection, post-monitoring, surveillance? All of which I'm a hundred percent in favor of. But if we aren't balancing it, we are going to develop a digital divide because if we put too strict criteria or reporting criteria or regulatory criteria, a critical access hospital will just say, "I can't do it. I literally can't do it." And thus, you have closed the door in helping out with my demand. And so, I just think we'll need to think, and this is probably based on a risk continuum and a few other things, but as we talk this afternoon, I think it's important to keep that context in mind and the practical realities of implementation here.

Dr. Elkin: If it's not too much trouble, I know the answer to his query and I think it would be helpful. So, there's been a move across the country to consolidate the health information exchanges, which are used in all the hospitals, including rural hospitals and healthcare facilities, and they roll up into statewide facilities. And there's actually a federal proposal that I think came out of the house this year to have all that data go to the National Library of Medicine and be a central repository for that kind of information. That would not add additional burden to any of the rural or community hospitals and yet would provide a large source of the kind of data that we would want as input to our process.

Dr. Bhatt: Thank you so much for that. I'll add two quick things, and then we're going to take a 10-minute break. The first is we've spent a lot of time between yesterday

and today talking specifically about existing devices, images, technologies, and the incorporation of GenAI into them. We spent a little bit of time talking about multimodality, and I just want to bring that back again for one more time because I think the way we think about it this afternoon will be different: multimodal inputs that have to do with clinical decision-making are in theory clinician facing. You asked this question yesterday, Troy, what about patient-facing versus clinician-facing? They're in theory just information for a clinician of whatever expertise level, and yet may have a greater impact on patient outcome than the interpretation of any type of image pathology, radiology, EKG. And so, I want us to keep that in mind when we talk in the afternoon.

The one other thing that I will say with my clinician hat on is the conversation started yesterday morning with Dr. Califf talking about fiscal advantages to the use of AI. And today in our discussion, the concept of the existing number of clinicians doing more work because AI makes them more efficient came up. So, for the sake of the clinical conversations I've heard before coming into the FDA today, I will say there is a fear for wellness among clinicians that they will be asked to do more because they are more efficient rather than thinking about making the system efficient rather than the individual. So, I just wanted to mention that with my clinician hat. Troy, let me give you the last word before a 10-minute break.

Mr. Tazbaz: I'm really glad you brought that up. And often I provide these statistics, and it's not to provide fear, but what is the opportunity or the lack of opportunity actually here and not doing something? And so, by 2030, we will have every single baby boomer retired or retirement eligible, which means that they will also be Medicare eligible. That's the demand side. The supply side is that we have 22 million people

working in healthcare right now. It's projected by 2030 that six and a half million of them are going to retire. And right now, we're projected to replace them by 2 million. So, this is not a healthcare question, this is an economic question. What happens when there's a supply demand imbalance, inflation? And you can apply inflation in this context, not just on a per capita cost of healthcare, which we are by far the most expensive country in the world. And Dr. Califf said, this is one of his key talking points, yet we don't see the outcomes. We're on the lower end of life expectancy. And to the second point of the inflation, time to get care. These are big problems, right? So, although this is really focused on AI, generative AI. AI is a tool, but we have to address this imbalance and so in whatever which way tools that we have at our disposal to be able to address them. And I think that's the bigger problem. It's not our artificial intelligence about job replacements. It is actually about this imbalance.

So, I welcome additional conversations because this is fascinating for all of us involved. And at FDA, we do actually think about these issues and we haven't really come up with a clear kind of approach towards tackling this different paradigm. And a couple people use the word paradigm because this is a different paradigm than what we're used to in dealing with medical devices. So, I appreciate the conversation.

Dr. Bhatt: Thank you so much. Let's take a 10-minute break, and then we'll solve the world's problems. Thank you.

Sub-Topic: Postmarket Performance Monitoring

FDA perspective – Approaches for managing changes for GenAI-enabled devices

Dr. Bhatt: Thank you all so much. Our 10 minutes are up. It went by a little too fast. I'll take a little volume. Oh, yeah. Thank you, guys, perfect. If everyone could please

take a seat. You guys should just touch your phones to one another. Don't type it in, just touch them and it'll magically-- Yeah, no. Okay. Maybe that'll take longer. Yeah. I am inviting my Panelists back. I'm sorry. That was a quick bio break. Apologize for rushing everybody. Can somebody who's near the door poke their head out and just ask for Committee members to come back in? Great.

We will now proceed with today's agenda. We'll hear five presentations, after which the Panel will have the chance to ask questions to the presenters. We will start with the FDA's presentation from Jessica Paulsen, who is the Associate Director for Digital Health in the Office of Product Evaluation and Quality at CDRH. Ms. Paulsen, you may come up and begin. Thank you.

Ms. Paulsen: All right. Good morning. So, I'm Jessica Paulsen. I'm the Associate Director for Digital Health in our Office of Product Evaluation and Quality at FDA and I'm excited to be here to chat with you about some approaches for managing changes for AI-enabled devices. So, we're going to start by again, underlying the importance of a total product lifecycle or TPLC approach. And then I want to focus on two different approaches for postmarket monitoring, including special controls and predetermined change control plans or PCCPs, which you've heard a little bit about, but we're going to do a crash course together.

FDA has long promoted a total product lifecycle approach to regulating medical devices, including AI-enabled devices. You know, this has become increasingly relevant and important for medical devices that are incorporating technologies that are intended to iterate or change at a much faster pace than ever before throughout their lifecycle. So, this really helps us push our needs to consider how manufacturers can

effectively monitor the postmarket performance of their AI-enabled devices to make sure that these devices continue to be safe and effective throughout their lifecycle as they change.

So, very briefly, I did want to mention that we have worked collaboratively within the International Harmonization and Regulator Community focused on establishing good machine learning practice guiding principles. These aim really to facilitate the development and assessment of high-quality AI/ML-enabled technologies. And I mentioned these just to, again, underline the importance of a total product lifecycle approach, which is listed in the guiding principles as well as the need for good performance monitoring of these technologies.

And I did want to take a moment also to just share about the current landscape. The number of FDA authorized AI-enabled devices continues to increase. We update our list periodically. This list is maintained on our website. We are at as of the last update this past summer 950 AI-enabled devices that have been authorized by FDA. And I did want to make it very clear we have not authorized any generative AI-enabled devices, nor have we authorized any adaptive AI-enabled devices.

That said, I thought it would be helpful to share an example of a recent authorization for an AI/ML-enabled device. This is the Sepsis ImmunoScore. So, this software is intended to identify patients at risk for having or developing sepsis. So, the way this software was designed is that it uses data from a patient's Electronic Health Record in conjunction with other lab tests and clinical assessments to really aid in the risk assessment for the presence of or progression to sepsis.

Given this intended use and the nature of this design or this device, it's important that it continues to perform as intended throughout the lifecycle and in the real-world setting. So, as part of the De Novo process, and we do this for all De Novos, we evaluate the risks associated with that product and to mitigate those risks, we establish special controls or requirements for that device type to again mitigate those risks. Thinking about this example specifically, a couple of the risks that came up during that review are the risk of bias as well as risk of poor quality or missing input data for that device type.

So, the new regulation actually included a special control that requires manufacturers to develop and implement a postmarket performance management plan. That is just one of the special controls, but again, given the focus of this today, I wanted to highlight that one because I think it could be pretty helpful as we think about the regulatory approaches to generative AI-enabled devices.

Taking that example and kind of zooming out a little bit, I did just want to remind folks that we consider intended use and technological characteristics within our broader risk-based regulatory framework. So, when thinking about generative AI, there are technological characteristics of generative AI that may sometimes introduce new or different risks for a particular generative AI-enabled product, which raises in turn new questions of safety and effectiveness. What that means is that a De Novo may be necessary to regulate that product and through that De Novo process, the output is that FDA would then establish a new device type, new regulation, and that would be accompanied by special controls as appropriate for that new device type.

FDA may require special controls that are unique to generative AI-enabled devices when needed to provide that reasonable assurance of safety and effectiveness of the device. Such special controls could include requirements for postmarket monitoring, such as the need for a performance monitoring plan. Again, this is why I mentioned that example, just to give you an example, it could be a performance monitoring plan like that one, but obviously there's an open question around what would a good or appropriate performance monitoring plan look like for a generative AI-enabled device? And that is where I think some of your input later today would be helpful.

All right, so shifting gears, I want to talk about Predetermined Change Control Plans for Devices. So, this again is another mechanism that I think could be particularly useful as we think about regulatory approaches for these technologies. So, by way of background, Section 515C was added to the Food Drug and Cosmetic Act in 2022, and this really clarified FDA's authority to authorize devices with PCCPs which include planned changes to devices that would otherwise require submission. So, it may be said a little differently, this is a least burdensome mechanism for both FDA and industry to allow a company to pre-specify how they intend for their device to change and FDA review that and authorize that such that then they wouldn't have to come back to us to make those changes. They'd be able to proceed with those modifications.

So, the scope-- This is not in any way limited to software or AIML. It is broadly applicable to all medical devices, and it is in effect right now, it is self-executing, which means that today manufacturers may submit an FDA may approve or clear a PCCP for a device at this time. And we've been doing just that, we've authorized PCCPs for devices across the spectrum in 510(k) De Novo and PMA.

We do have a draft guidance that we issued focused on providing recommendations for PCCPs for AI/ML-enabled devices. This guidance really outlines our science-based approach to putting safe and effective advancements in the hands of users and healthcare providers faster. Again, our goal here is to help increase the pace of medical device innovation, and our approach as outlined has been informed heavily by our experience in regulating AI/ML-enabled devices to date.

Briefly, I just want to go through the components of a PCCP and these are described in detail in the draft guidance. First, we have the description of the modifications, then we have the modification protocol, and then finally we have the impact assessment. So, together, all three of those are what we call a PCCPs.

First, we have the description of the modifications, which is exactly how it sounds. This is where a manufacturer would provide a detailed description of each of the planned modifications that they intend to make to their device. So, this includes very specific modifications that can be verified and validated and that those modifications are provided at a level of detail that actually allows FDA to do our assessment so that we can reach our conclusion on that device, that it's going to be safe and effective with those changes. And the guidance talks about how we recommend that manufacturers specify whether those plan modifications would be implemented manually or automatically. We also recommend that manufacturers specify if proposed modifications would be implemented globally or locally.

When thinking about the modifications that are appropriate for inclusion in a PCCP, those modifications that are really intended to maintain or improve safety or

effectiveness of the device, they are specific, they're able to be verified and validated, and in accordance with 515C, they must maintain the device within its intended use.

All right, so now that we understand the changes that are being proposed, what we would also see is the modification protocol. This is where a manufacturer describes their methods that would be followed when developing, validating, and implementing the modifications outlined in the PCCP. This includes the verification and validation activities. It includes predefined acceptance criteria, and this is what would support that. Those modifications can be made to ensure that the device remains safe and effective, and the methods that are described within that protocol should be consistent with and support those modifications. It's also really helpful when manufacturers describe for us how those methods may be similar to or different from methods that are used elsewhere in that submission or maybe a previous submission for a different version of the device.

In the draft guidance, we go into some detail around components of the modification protocol, again, specific to AI-enabled devices. So, we focus on the importance of a manufacturer describing to us their data management practices. For example, how are they going to assure quality, adequate quality of the data? How are they going to determine a reference standard? For retraining, what are their objectives and how will they implement that retraining? For performance evaluation, we're looking for assessment metrics, we are looking for statistical analysis plans as well as performance targets. And then for update procedures, and again, I know this has been a common theme today, it's really important that we understand how and when updates will be implemented and how users will know that those updates have happened because transparency again is incredibly important here. And then also a device monitoring plan could be a part of update procedures.

And then the final component of a PCCP is the impact assessment. This is where a manufacturer documents essentially their benefit risk assessment for the PCCP such that FDA at the end of this review can determine that the risks have been adequately identified and mitigated such that the modifications within the PCCP could be made, and the device would remain safe and effective.

We do also recommend that labeling be updated in accordance with the PCCP, again to make users aware of the modifications that have been implemented via that PCCP. This includes providing a description of the implemented modifications that would include a summary of the current device performance, a description of how modifications were implemented, as well as a description of how users will be informed of the implemented modifications.

And as you may be aware, many of our FDA authorized devices include public decision summaries. So, we do recommend that the public facing documentation include a summary of some of the key PCCP information, again to support transparency and safe and effective use of the device.

All right, so we've made it through our crash course on PCCP. I did want to leave you with some considerations as we think about how we can leverage PCCPs for adaptive algorithms. This is not an exhaustive list by any means, so please don't take it that way. But really just a few thoughts that are top of mind as we think about how we might leverage PCCPs here.

You heard me say a couple of times the importance of having specificity when it comes to how exactly the device will change related to the modifications. I think we have to think about and appropriately calibrate, you know, how specific can the

modifications be for an adaptive algorithm in a PCCP? Would that look different when thinking about a locked algorithm, for example? What boundaries or guardrails could be established in the PCCP to better define the range of automatic updates, for example? How are we going to think about postmarket performance being monitored over time to assure the device's performance is maintained or improved? And how do we make sure we account for monitoring across multiple sites, allowing for local and sites specific adaptations? How will labeling again be updated when modifications are implemented automatically to inform users? We want users to have the right access to the right information at the right time. Appropriate notification requirements could be considered here. This is again tying back to 515C where notification requirements would be needed if the device does not function as intended pursuant with the PCCP.

So, again, with that I will wrap up, but I just want you to keep in mind, as you all are thinking today and you're faced with FDA's questions, these are a couple of regulatory approaches that could be helpful as we think about how best to regulate these types of technologies. With that, I'll thank you for your time.

Stakeholder Perspective – Supporting Health AI for Impact

Dr. Bhatt: Thank you so much. Our next presentation is from Dr. Christopher Longhurst. He's a Clinical Professor of Pediatrics and Biomedical Informatics, Chief Clinical Innovation Officer, Associate Dean at UC San Diego, and we can play video presentation now.

Dr. Longhurst: Hello, colleagues on the FDA Digital Health Advisory Committee. It's a pleasure to be here with you virtually today, and I appreciate the opportunity to discuss some thoughts around regulatory oversight and the total product lifestyle for generative

AI devices and algorithms. My name is Chris Longhurst and I serve as the Chief Clinical and Innovation Officer at UC San Diego Health, as well as the Executive Director for the Jacobs Center for Health Innovation at UC San Diego. And I want to share with you off the bat three key ideas.

First of all, I think the FDA should reconsider the 510(k) pathway for AI algorithms and should be empowered with the increased authority to require transparent outcomes data about products. Number two, I think that you should have increased authority for postmarketing surveillance. And then number three, I really think this is a shared responsibility of health delivery organizations and I think we as hospitals and health systems should be subject to a regulatory requirement for local testing and monitoring that can help you with that postmarketing surveillance.

So, I'd like to share three quick case studies and examples of how this might work. I've been doing a lot of thinking recently about AI and patient safety, particularly since we're celebrating the 25-year anniversary of the Institute of Medicine report *To Err is Human*, and yet we know we haven't made the progress we wanted to make in the last 25 years.

Recently, I identified a dozen different areas in which AI can help us with patient safety. The first case study I'll share is about chest X-rays. At UC San Diego Health the pandemic started about a month early because we got a plane full of expatriates from Wuhan here at the Miramar Air Force Base. And by the time that end of March rolled around, we actually had deployed an AI algorithm to help with detection of COVID pneumonia. We had some anecdotal data that this was helping, but we undertook a rigorous study to really identify whether it was achieving impactful

outcomes. And what we found was that over the course of the summer of 2020, that one in five patients who got a chest X-ray had a different clinical decision because of the AI.

This is a really remarkable outcome, and in most cases, what it was about was deciding whether or not to test. So, you can imagine that as testing became ubiquitous in the fall of 2020, that this algorithm actually became less useful and in fact, concomitant with the AI bill of rights that the president later published that AI should be monitored for outcomes on ongoing basis rather than a single offsite evaluation, we actually chose to discontinue this AI-based alert by the end of 2020.

That's one important learning. The second was that a couple of years later, this article was cited in a review article about AI and COVID. Turns out there were over 9,000 articles published on this topic in 24 months, and this group found just four that had clinical outcomes. You can imagine my surprise when one of the two from the United States was the one I just shared with you. So, this really identifies and highlights a big issue, which is the health AI paradox. There are many, many, many algorithms being proposed and published, but very, very few that have actually been evaluated for the impact on clinical outcomes. We're going to talk a little bit more about that.

And the second example is around sepsis. I don't have to tell this audience that it's a dangerous condition that affects 1.7 million adults on an annual basis and a third of a million Americans die on an annual basis. So, it's something that we've been tackling locally at UC San Diego Health with both traditional methods and novel AI algorithms. And credit to Dr. Gabe Wardi and Shamim Nemati pictured here who published earlier this year that their deep learning sepsis prediction model actually had a significant impact on mortality. In fact, the 17% is a combination of our two emergency

departments and the one in Hillcrest which serves a more underserved population actually had a bigger impact in mortality. Not only are we having meaningful clinical outcomes, we're actually closing health disparities with this AI algorithm as well. And this paper was highlighted by the editor of the journal who wrote an editorial really highlighting the importance of these types of meaningful outcome metrics and suggests that funding agencies and journals alive should be shifting to prioritize these types of studies.

Now, in contrast, there was a first ever FDA clearance of a sepsis AI algorithm just a few months back that was based off of the 513(f)(2)(De Novo) process. And going through the 50 pages of data that's publicly available for it, what you find is that it's a retrospective analysis of data with no meaningful clinical outcomes. And so, I hope that this works, but we don't really know. In fact, it's kind of like approving a drug and saying that this new breast cancer drug has very few side effects, but we don't have any idea about its efficacy. In fact, some of my colleagues here at UC San Diego recently proposed a regulatory strategy for AI that's outcome centric and would require companies to demonstrate AI tools produce clinically important differences in patient outcomes before being brought to market.

That takes us to the third and final case study, which is something that came up just earlier this year when we started tackling generative AI in the Electronic Health Record. Now, our AI principles at UC San Diego are listed on the left. I shared this story with the White House Task Force last fall, and that was encompassed in the statement that was released in December.

We were one of the first two health systems in the country to implement generative AI and Electronic Health Record to help respond to patient messages. And one of the principles we adhered to was accountability. So, none of these messages go out without being reviewed by a human, which is why there's only two buttons to start with a draft or start with the blank reply. There's no button that says, "Just send now". But the second principle we talked about was transparency, and the question was asked whether we should be transparent with the patients that we serve. And some of our doctors even argued we use dot phrases and macros today, but we don't have an addendum suggesting that. But our patient representatives really highlighted that AI can be scary for patients. And so, anytime one of our clinicians uses that start with a draft button, an addendum is automatically appended that transparently shares with patients that part of the message was generated automatically before being reviewed and edited by their doctor.

This actually came up recently in the New York Times article where you see Teddy Rosenbluth's subheadline saying, many patients have no clue that replies are software generated. In fact, in California, one of our lawmakers, Lisa Calderone, recently proposed a bill that was signed by our governor into law Assembly Bill 3030 that's going to require disclosure to patients and families when AI is used to generate communications, whether they're written or telephonic.

So, where do we go from here? Well, again, number one, I think what really matters is outcomes. And we recently called along with Aneesh Chopra, the former White House CTO and others for a network of AI implementation science centers. What we don't need is assurance labs looking at algorithm performance because the real-world impact on outcomes can't be predicted from performance, just like in that case of

sepsis, where we actually sacrificed accuracy of the algorithm in order to put it earlier in the workflow where it could have more impact for our clinicians' decision-making.

The second thing is we can't look at the hospitals and health systems off the hook. This is a shared responsibility. And so, we called along with Rajeev Wani and other co-authors for CMS conditions of participation to require AI oversight when deployed in health systems.

So, again, to recap, I think Dr. Rob Califf agrees with this concept that health systems have to step up on AI regulation and healthcare delivery organizations should be held accountable. Along with that, we think the FDA should reconsider the 510(k) pathway for algorithms and should have increased authority to require transparent outcomes data and increased authority for postmarketing surveillance. So, again, thank you for your time and look forward to the vigorous conversation. Take care.

Stakeholder Perspective – Technological Innovation and Considerations for Postmarket Performance Monitoring within Radiology

Dr. Bhatt: Wonderful to have Dr. Longhurst join us, albeit virtually. Next, we will hear a stakeholder perspective from Dr. Nina Kottler. Dr. Nina Kottler is the Associate Chief Medical Officer for Clinical Artificial Intelligence at Radiology Partners. Dr. Kottler, you may begin.

Dr. Kottler: Thank you, and thanks for having me. I am enjoying this robust discussion sitting here watching both days and trying not to chime in yet until it's my turn, but I am noticing that from the clinical stakeholders, the people that are using this tool, which is mostly in radiology but also in other arenas, we seem to be saying a similar message. And I'm going to be echoing that message today.

I'm going to talk to you today about postmarket performance monitoring in radiology because that's where we've started and that's where we've been doing it. And just to give you a little background, I am the Associate Chief Medical Officer within Radiology Partners. Radiology Partners is the largest radiology practice in the US. We are onsite in 35 states, and we have a sister practice that is a teleradiology practice that performs evaluations remotely in all 50 states. We have a very massive and variable footprint from very rural sites to level one trauma centers and urban environments to academic groups where we teach residents about radiology and everything in between. So, the variability in what we see is quite massive, and it's been very instructive to us.

We also have a very big footprint in terms of scale. We do 10% of all of the imaging that is done in the United States, it's interpreted at our practice. And we've been using artificial intelligence since 2016. So, we have a massive amount of deployment experience. In fact, when I looked back, this was data from May. When I looked back at how many exams or reports we have processed, either through computer vision or a language model, it was 200 million, and that was back in May. Now, when I look at that process, which is accelerated just by the end of this year, we'll have processed a hundred million exams through some kind of artificial intelligence system or reports. So, it's that experience that I hope to bring to you today. Here are my disclosures.

The first question is how is artificial intelligence really the new generative AI? What you guys are asking about being used in radiology today. And there's four ways that I'm seeing the vast majority is having an output of some kind of text. So, it's usually a textual output. Either it's text input and text output or it's text and or a computer vision, like an imaging input to get a text output. We do have some imaging

outputs that's not nearly as common. It's going to get more and more common. And I'm not going to talk about that specifically today. However, I think a lot of the principles that I'm going to speak about will apply.

I will mention that continuous validation in artificial intelligence, whether it is the new generative AI models or the old-fashioned models, is something that is evolving. There isn't a best practice out there that everyone is using, but there is a lot of good theory, and we're going to be hearing and seeing a lot more and more of that over the year.

Today we're going to talk about two things. I'm going to start with a monitoring that we have done on a GenAI system, a GenAI system that's used across 40% of the radiology practices that are out there, and so it's significantly in clinical use. We wanted to take a look at how accurate that was in the clinical environment. And then the second thing we'll talk about is, you asked, what could be a framework for how we do this continuous monitoring with generative AI in the future. So, we'll go through that second.

The reason why I want to talk about this first section is because I want us to be a little more precise about our language. All right, so this is a radiology report. The radiology report has an initial section that I'm calling the preamble. It has details about the patient and the reason for the exam, the type of exam that was done, and then the radiologist will dictate something in the finding section, that is: What are we visualizing in this image? We articulate that in the finding section. And then the impression is exactly that. It's our impression of the things that we saw and what that means for that patient. And as the consultant for the other physicians, that's what the radiologist's job

is, the physician's physician. That impression section is: How do we think that you guys should think about? How can we help you as the referring clinician take care of your patient better?

We have a generative AI system. Again, it's being used quite significantly throughout the radiology practices all across the US. And that generative AI system will take the findings that the radiologist dictates and when they get into that section of the report, it'll automatically populate that for the radiologist. Now, we don't send this off. This is not something that is under FDA guidance, but we still don't send it off. It's not necessarily something that we think is safe to send off. We have a radiologist right there in charge of the radiology report. They take a look at that impression, they read through that, and they may change it. They may edit a part of it, they may edit all of it, or they may edit none of it.

So, what we did is we wanted to compare what was the generative AI output to what the final edited report or unedited report was. We looked at 3000 radiology reports from across our very highly variable practice, and we compared those two together. Now, when we compared it, we didn't just compare it with a large language model and expect that the radiology report was accurate. There are people here that have mentioned a few times, everyone says different things. So, we had a Panel of experts of radiologists that reviewed these, and we wanted to validate the accuracy, but we also wanted to determine if the changes that were made were clinically significant.

And I'll go straight to the results of those. The result is that the AI, the generative AI, had a 4.8% clinically significant error rate in generating the impressions. That's one out of every 21 reports would've had something that was clinically

significant, meaning would've affected the patient. Now, what's super interesting about it is we looked at the final result. Now, the final result actually only had a 1% error rate. And why is that?

We decreased the error rate significantly because we had the radiologists, who are the experts, we had them in the loop evaluating the AI. So, it's not just a human in the loop, which I absolutely agree with. It's a specific human in the loop. It's an expert human-in-the-loop. So, we had an expert human in the loop. The 1% error rate is significantly decreased from the 4.8%, but that doesn't tell the whole story.

The actual error rate that was decreased, it decreased to 0.5% by the radiologist. So, you're saying, "Well Nina, you're telling me that it was 0.5, but here you have that it was a 1% error rate." And that's because when the radiologist re-dictated that change, because they're not generally typing it, they're dictating, they introduced another 0.5% error. So, the actual decrease in error rate went from 4.8% to 0.5% with an expert human in-the-loop. That's a 90% improvement in error. And this is something that we found, whether it is in pre-validation before we deploy an AI or in this kind of monitoring after an AI is deployed, that there's significant mitigation of the errors that the AI makes at the point of care, especially when it is an expert in-the-loop. And especially when you educate those experts about the types of errors that you tend to find.

It's very hard to go to a radiologist and say, "Just be careful on every single exam because 5% of the time it's going to make a mistake." How do you find that five out of a hundred cases? I mean, it's like finding a needle in the haystack. The goal is you either make the haystack smaller or the needle bigger and the way you do that is

you try to understand the types of errors that the AI is making. And this is what I think is the most important part for us, whether it's the FDA or the market to do in advance of deploying these models, is understand where it's making those errors so that we can educate our end user.

So, a lot of times we are very generic in our language and when we talk about the error types made by generative AI, we call them all hallucinations or confabulations, but that is not specific enough to bring as education to the physicians for them to understand where the errors are going to be made. So, I broke them down a little bit more for you and I'm going to go through examples. Numerosity errors, laterality or body site, recommendation, missed findings, other hallucinations. Let's go through them one by one and I'm going to show you exactly where we found those errors and what they looked like. Now, this is important because if we use the right language, language helps us with safety. You first have to have the language, the ontology to describe it. Just like we do for CT and MRI artifacts, we see them all the time, but we still deploy CTs and MRIs even though they have artifacts, but we have a language for them. We educate our physicians about what they are so that they can mitigate and understand them at the point of care.

Let's start at the top with numerosity errors. Those are errors of number, and every time I show you these images it's going to first show you the finding section so you can see what the AI was using to make its decision. Then I'll show you the generative AI impression. That's the summarized impression from those findings. And then last, I'll show you the final impression, which is the edited impression by the expert in-the-loop. And you can see in this case that there's a lot of numbers that were dictated. Now, this is a very small part of a CTA examination where lots of numbers are

dictated. I just took a segment of this; this is not the full report. And you can see that there's three numbers that were dictated 70%, 30%, 40%, and they were all about a stenosis of a different vessel.

The generative AI was confused with those numbers and didn't bring them precisely into the impression. In fact, it said 70, 40, 40 instead of 70, 30, 40, and that is clinically significant. Now, it could have said 30, it could have said 300, it could have said something else. And this is the thing that we have to be careful about. So, how do you make sure that your radiologists are seeing this? Well, they have to know that more often the AI is going to generate errors when it comes to numbers. And here you can see they corrected it.

Here's another numerosity example, and this is one that we see commonly. The dates when they're dictated separately, often get combined together. That has a clinically significant output when you're looking at something longitudinally. Here it took the 10 and the 25 from the first date and combined it with the year 2019 on the second date. The radiologist expert in-the-loop corrected it in the final impression.

Let's go on to the next one, which is laterality. Now, laterality is one form of this error that we're calling a "fill in the blank" error. Why do we call it the "fill in the blank" error? It's something that I think we can understand as clinicians. It means that if you're not dictating something that is normally dictated, the AI is just going to guess and fill in that blank that you left. So, in this example, the initial findings are talking about the femur. Now, we have two femurs. There's a left femur and right, you don't just say "The leg has this finding," you say the left leg or the right leg. Now, the exam is usually on one leg, but generally we dictate this. So, the GenAI didn't know what leg it

was, but it's used to having a laterality associated when you say femur and so it guessed. Now, it guessed wrong in this case. It allowed the radiologist-- The radiologist found that allowed them to change it. Actually, also allowed them to find that error in the findings and fix the error in the findings. This is also a benefit of the AI, what we call a mirror effect as it mirrors the errors that are in your findings, brings them down to your impression, and actually 1.7% of the time the radiologist is able to find errors in their findings because of this. And that's higher than the 1% error rate overall that this is brought down to.

It's not just with laterality. This "fill in the blank" error happens to anything that we're saying consistently. So, body site, stability of something over time, severity, size, and the clinical indication. We're seeing this quite frequently.

The next one is recommendation. Now, we've actually turned off the recommendation function of this GenAI. It has a separate function that will look at population health guidelines and provide that population health guidelines. We've turned that off because we have another tool that does it, but there's still recommendations built in the embedding space because there's recommendation in the radiology reports and sometimes it'll pull in recommendations that aren't appropriate. So, we tell all of our radiologists, anytime you see a recommendation, you have to be more vigilant because it may be making an error. And in this case, it is, it's recommending an MRI for a benign finding. It even says "non-aggressive." In the final impression, the radiologist was able to change that saying this is favored to represent something that is benign and doesn't require this MRI follow up.

Missed finding. There are times where the AI does not bring the relevant finding down into the impression section. Maybe it's not built in the embedding space and seen it enough. And in this case, it does make a difference. So, this was clinically significant. This is a urologic examination where there was distension of the kidneys and usually that's related to some kind of obstruction, and you're looking for the obstruction. They didn't find an obstruction, but what they did see is that the bladder is really distended and that can be a cause of hydronephrosis. The impression brought by the AI did not mention the bladder distension, which is very relevant in this case. The radiologist added it. So, a missed finding.

And then other hallucinations. I'm going to break down a little bit more for you. So, this one we call the nearest neighbor and this is related to the embedding space. What kind of things are commonly said together in your radiology review reports for which the embedding space has associated those?

In this case the initial finding is a small amount of free fluid in the pelvis. One of the most common things, especially in women, the cause of this is just ovarian cyst rupture. We see it all the time. So, the GenAI associated those things together. Unfortunately, this is a male patient, so not quite appropriate, but the radiologist was able to change that and that is something that we call a nearest neighbor effect. Here's another kind of hallucination where it's contradicting itself. In this case there was a contrast enema that was done to look for a connection between the rectum and the vagina. And they said they did that, but they didn't find this fistula. But the GenAI impression highlighted at the bottom said, "Well, they didn't do a contrast enema." We actually did, that's how you figured out the results. The radiologist removed that. That's a contradiction to the initial statement.

Another case is one of redundancy and in this examination-- We find this frequently, like you're dictating a CT scan of the chest, there's often more than one lung nodule and we'll mention a bunch of them, maybe not all of them. We'll say there's a whole bunch and here's at least the top five, but there'll be multiple. And you can see there's a four millimeter and a six-millimeter nodule in this case. It mentioned both of those in the first impression, great first impression, but then it re mentioned the six-millimeter nodule again, and that can be quite confusing. We don't want to be overly treating and mentioning things twice so that we think of them as separate nodules when in fact they're not. So, the radiologists fix this.

All right, so hopefully you got an idea from our actual validation of using this in production with radiologists overseeing it. A few things. Number one, a better language, a more specific language that we can use to understand how to describe these findings because you need the language in order to educate and you need education to mitigate at the point of care, which absolutely can happen.

Now, let's talk about a framework for continuous monitoring. How could we potentially do this at scale? Now this is something that Keith mentioned and a lot of other people have talked about. The way we're doing this now is we have our data, whether that is imaging data or a document that goes into an AI system. The AI system does the evaluation and you get your AI results. We are then in radiology. Most of it is being done for a radiology report. So, we're comparing it against the radiology report and you're looking for concordance or discordance between those results.

Now, you can either do that at scale with no clinical adjudication, you could do it the way we did where we adjudicated every case. You have to probably do something

in between because adjudicating every case that you just don't have enough RADS to go around. But it is an important component to make sure you have that expert group that is looking to validate to make sure that there isn't automation bias and other things that you don't find automatically. But those are the first two ways that you could evaluate the performance of a generative AI at scale. One automated on its own, two with some component of a human in-the-loop.

That's not the only way. There's a third, fourth, fifth and sixth way. So, I'm going to show you a third one. The third one is called Ensembling. And what you do in that case is you create multiple AI models that are doing the same kind of task as your initial AI model. They can be much smaller; they can be much cheaper. They don't need to necessarily go through FDA clearance. All you're using them is for a quality assurance check. And you look to see if the output of this group of AI models, if one of them or two of them are different from the AI that you're looking at, then you could adjudicate those cases. And that's a way to see something before the radiologist actually gets the result.

Another way to look at your AI results continuously before the radiologist gets the results is to look at the data input. We all know that if the training data was not sufficient, then you're not necessarily going to get a great result. In this case you could look to see: is your data that you are using for that case in or out of distribution from your training data? I will say it is far more effective doing it this way than it is for me at the point of care to have a nutrition label about what was seen, the training information done by the AI, because I'm not going to be able to use that at the point of care. But I can use this at the point of care where if something is looking and telling me this is out of distribution from the training data, that helps me at the point of care be more vigilant.

Alright, so here are those four plus two more which people have mentioned. There's also model stress testing that you can do and of course user feedback. User feedback is essential. You're not going to get it from everyone, so it'll be a little bit biased. Now, each one of those will have strengths and weaknesses. And the point here is that there is no one way that you all have to do this. There are multiple ways that we could do this and depending on the risk of the device and who is using that, if it's an expert or a non-expert, all of these things can be combined using either one of them or multiple of them depending on the model that you're evaluating and the model risk.

So, today we talked a little bit about how generative AI is being used in healthcare. It should be monitored, it absolutely can be monitored, and that will provide us far better information than we're getting now. Pre-validation. What I do think we should be doing pre-validation is understanding the types of errors that are made. That's the risk. And using that to educate the radiologist or the end user at the point of care. Our validation on this GenAI model suggests that it should not be autonomous, it can be used-- It shouldn't be used on its own. It should be used with an expert in-the-loop. But that is definitely useful and the experts should be educated on the types of errors that are made. Thank you.

Dr. Bhatt: Thank you so much, Dr. Kottler. Next, we have Dr. Grace Cordovano. She's a Board-Certified Patient Advocate (BCPA), founder of Enlightening Results and Co-Founder of Unblock Health. Dr. Cordovano, you may proceed.

Stakeholder Perspective – In Real Life: The Patient Health Information Journey and Generative AI-Enabled Tech Impact

Dr. Cordovano: It's an absolute honor to have the opportunity to share the patient and care partner perspective now on postmarket performance monitoring of GenAI-enabled medical devices. As a Board-Certified Patient Advocate, my day-to-day is working with patients and families from the point of diagnosis through survivorship or end-of-life care planning. Many of my patients rely heavily on different types of modalities of imaging as well as pathology, have numerous specialists in the mix and often are faced with making very difficult life altering, life limiting decisions in sometimes less than 60 seconds. Access to credible information is key. In addition to my professional advocacy, I'm the primary care partner of two disabled adults and I'm a patient myself. This topic deeply resonates with me and it matters a lot.

So, as I dreamed big about what comments I could prepare today to prepare something and a conversation that would shape and set the table, I wanted to start with some statistics from a recent community-based survey that was published in September conducted by the Light Collective. They had a community-based survey of 377 participants who self-identified as diverse patients, care partners and advocates. And some highlights from this survey. 91% of respondents stated that they are concerned and want to be informed if AI is being used to make decisions about their care or if it's being used about communications of their care. 74% of patients and caregivers and advocates that responded stated that they're moderately or extremely concerned about potential unauthorized disclosures of their information. Privacy matters. 97% of respondents stated that they were at least somewhat concerned about potential negative impacts of AI and harms that could happen with AI being applied to their care.

When we hear stats like this, of course we clearly need to get postmarket monitoring right as soon as possible. So, dream big with me as I look at a different approach, a patient-centric approach to postmarket performance monitoring. And as I thought about that, four arenas really crystallized for me. The first one being partnership and co-creation with patients. This meeting was divided in pre and postmarket considerations, but the reality is from the patient perspective, robust postmarket performance monitoring will require intentional partnership with diverse populations, diverse patient voices included throughout the process in a co-creation process that goes throughout the entire product lifecycle.

If I had to summarize, I would say patient communities feel very strongly that oversight of AI has to be democratic and there have to be checks and balances on power. From the patient perspective, while safety evaluation of AI of course needs to be deployed and evaluation of AI needs to be implemented by industry, from the patient perspective, industry can't be the only one driving the bus and the only one in charge. The patient stakeholder perspective matters and needs to be included continuously and throughout.

We've heard a lot about outputs. Let's talk a little bit more about outputs. When an output is generated by one of these AI powered medical devices, patients don't see them. I loved the presentation right before me, which dove into some of the intricacies of the outputs. But the reality is if an output that's generated is used to make decisions about an individual's care or coordination of care, a patient should have access to it. It should be recognized as part of the designated record set. It should be part of our medical record. This is essential for trust and transparency's sake. These outputs were

going to require standards for exchange if they become part of our medical record. These outputs matter for informed decision making and shared decision making.

I'll share a personal story. I have a family history of breast cancer. I've been receiving mammograms for the last 14 years, and I was excited when my last mammogram came in and I saw that it was mentioned that there was an enhanced breast cancer detection software applied to my mammogram. The mammogram was normal. However, there were a number of benign findings. So, here for 13 years, nothing. Now I've got [Makes a sound with her voice.] Anyone who knows how I operate. I immediately called the imaging center and I asked for a copy of the output. I wanted details. Where were these findings coming from? Could these be false positives? I wanted more details. What did my other mammogram images look like? Could that AI be applied to those and are there trends? Well, needless to say, they didn't know what to do with me. I got a "Ma'am, we don't do that." I asked if I could have a copy of my mammogram without the AI applied so I could do a side-by-side comparison, they escalated me to a manager, to a supervisor. We can't accept the "Ma'am, we don't do that" because that is patient engagement, and it mattered to me because I'm at a higher risk for breast cancer.

Even though I was told everything was benign, a month later I got a confusing letter saying, "Based on your mammogram, you now need to go for an MRI." So, there's just discrepancies and I care and I'm digging in and I know that there's other patients and families that are digging in as well. Now we're going to do co-creation, we are going to partner with patients. We're going to look into getting access to these outputs that are so meaningful and filled with rich information. Now, patients really need transparency on where GenAI is being used in our care.

I outlined my comments for today and I started to get frustrated because how can patients-- I think it's wonderful that the patient voice is being included, but we're at a complete disadvantage here. We have no idea where it's being applied in our care. We don't know at what point, who's doing it. And I actually have a list of questions that started to come through because I posted that I was presenting today. So, bear with me. Patient communities want to know about GenAI-enabled devices medical-- GenAI-enabled medical devices. Who is, when and where are these devices being used? How effective are these devices? Where are they being used? What training data was being used? What are the outcomes? What should be celebrated? Where is potential bias, discrimination and harm happening? Is this pinpointed anywhere? I want to avoid these places of care. The end. When are outputs not used and ignored? Did the doctor override it? How is that documented? What if that override was wrong? What are some GenAI-enabled medical devices? Are some better than others in safeguarding my patient information? What if an output leads to something discriminatory and harm happens?

As you can see, this is just a sampling of high-level questions that patient communities are asking, and it really highlights the need for patients to be included. And I'll plug my 2 cents in. It sounds like a great opportunity for a patient task force or work group.

Number four, patient as human in-the-loop. We've heard a lot about humans in-the-loop, but we don't recognize the patient as an active participant. Troy, I loved your comments about calling out the shortages that we're going to have in a few years, and I'm going to channel one of my mentors, Dave deBronkart, e-Patient Dave, who says,

"Let patients help." So, we talk about having humans in-the-loop with respect to safe and responsible use of GenAI-enabled medical devices.

Number one, patients must be recognized as end users of GenAI-enabled medical devices. Number two, we must be recognized as essential humans in-the-loop locally, regionally, and nationally and have the opportunity to provide structured feedback somewhere when it's applied to our care. So, we're warned about hallucinations. "Don't use it, don't use it. It's all these things." The fear mongering. But what do we do if we do use it and we find something wrong? We have nowhere to go. I'll close with another example. Not only have we not been offered a mitigation strategy, we don't even have a collaborative strategic approach either.

When I attend appointments with patients, let's say I'm in the emergency room and I have a patient that's been there for 36 hours waiting for a bed, I can go and complain to a social worker who can then maybe escalate me to the charge nurse, who can then get me to the office of patient experience and we can move things along, and escalate if this is an insurance issue. If I have a problem getting a copy of my medical records, I can go to the Office of Health Information Management. I can escalate to the Privacy Officer. If that doesn't work, I can file an information blocking complaint.

If I have a problem with a GenAI-enabled tool or any type of tool that's part of my care, my loved one's care or any of the patients I serve, who do I go to? Boots on the ground, there's no point person. So, yes, there's going to also need to be, when we think about postmarket monitoring, how can patients and families and advocates help and have either a person locally or maybe it's nationally, maybe it's a separate office as we think big about strategies. That's sort of my closing remarks to leave with you today,

and those are the four categories that I see necessary for building an infrastructure of patient-centric postmarket performance monitoring. Thank you so much.

Dr. Bhatt: Dr. Cordovano, thank you so much. Our last presenter of the morning is Dale Webster, who is the Director of Health AI Research at Google. Dr. Webster. Oh, Dr. Webster is on the video.

Stakeholder Perspective – Industry View on Effective Evaluation Methods and Post Market Performance for Health-Generative AI Products

Dr. Webster: My name's Dale Webster. I'm a Research Director at Google focused on Health and AI, and I'll be talking to you today a little bit about our experience so far with generative AI and postmarket monitoring, and other model evaluation methods. We'll start out talking a bit about generative AI and maybe how it differs from previous versions of AI that we're more familiar with like predictive AI. Then I'll go a little bit into the AI lifecycle, how we think about it, how it's similar and different, and then finally I'll deep dive into model evaluation, which I think is where the really interesting parts of the differences between the two come into play.

So, first a little bit about generative AI. So, in order to understand that, let's go back to say 2016 or so when AI was first sort of becoming more visible in the public eye. The technology had reached a point where suddenly it could do things that we hadn't really imagined it would be able to do anytime soon. Things like tell the difference between the cat and a dog from a photo or maybe even the difference between specific species of cats and dogs in ways that most humans couldn't do. It was definitely a very interesting time. It got a lot of people interested in the space and the method of choice then was convolutional neural networks. Around this time, we saw a

lot of additional research including our group into what this might mean for healthcare. And so naturally the first areas we looked at were in medical imaging and we used sort of the same kinds of models the consumers were using to search for cats and dogs in their photos to start looking at things like, can you diagnose diabetic retinopathy or DME from a fundus photo? Many other kinds of imaging analysis of this sort.

And that process-- So, that was almost a decade ago now, and I think over the last eight or nine years, we've learned a lot about that space. We understand it pretty well. We understand how these models work, how we measure them, things like this, and we're starting to see traction in actually getting them into patient care, which is great. Of course, fast forward to about two years ago and suddenly large language models, which are this kind of newer kind of AI hit sort of that similar level of readiness that we saw with convolutional networks in 2016, and suddenly became very interesting and people started to think about how to incorporate these into products and into the world.

A little bit about how generative AI is different. Large language models in particular. Instead of starting with a photo, a set of pixels or a sort of piece of constrained size data, instead they start with texts, a series of words and letters. Those are fed into a model. It goes through typically a transformer architecture, which is kind of the equivalent of a CNN convolutional network, but quite a bit different. And from this transformer, the model tries to predict what the next correct word is. You can kind of think of it at some level as a very fancy auto complete. This is a fundamentally different kind of model. The underlying math at the basic level is similar, but the architecture is quite different. The inputs and the outputs are different, and we'll talk a little bit about the properties of different and what that means for its use.

That was about a year or two ago where we first started to see it move into the product direction. And the last few years we've been watching this develop and what we're seeing is this rapidly evolving system of new models being trained and measured every day, new model architecture ideas still under exploration, and there's a lot of engineering and work going on to try to get better at these trade-offs that we're seeing with the models around costs, speed and generalizability in general. People are trying to drive down costs, increase speed, increase generalizability so we can make these more useful for products. It's a very rapidly evolving space, and this is similar to convolutional networks in 2017, 2018. They went through sort of a similar process. So, I think we can expect to see the same pattern where something like eight or 10 years, we'll watch this evolve and we'll understand these things much better.

That's kind of on the technical side, but let's think about when these models hit the real world, what does that mean? What's an AI life cycle for generative AI look like? And this is a great diagram pulled from the FDA around the AI lifecycle. Looking at this, really the diagram doesn't change all that much. These are still the same steps we use in generative AI that we used in more traditional AI approaches. The details I think within each of these steps will be different in some of these steps. It's not that different in others, it's quite a bit different. For the rest of the talk, I want to focus on two of these areas: real world performance evaluation and verification and validation. These are the areas that I think have the most important differences and the differences that we should spend a lot of time figuring out in the next couple of years. And you'll notice that the common theme for these is they're both the parts of a lifecycle where you're doing model evaluation.

One of the ways we do evaluation is postmarket monitoring. It's definitely something I wanted to touch on specifically in this talk. This is a diagram we use for our postmarket monitoring, to describe our postmarket monitoring for our imaging systems where we have areas where the models are deployed. At each of these areas we are randomly sampling images and the grades that the model's giving for these images, we're sending them back to a central point where we have humans and other automated methods looking at these grades, evaluating whether they're correct or not. And then we have sort of dashboards and experimentation we can run to assess when maybe something's changing out in the world that we need to update our model to compensate for and things like that. And really this level of postmarket monitoring, I kind of expect to be the same for generated value. That basic process is quite similar, but once again, it's this evaluation part of the end that's going to be tricky. We'll deep dive into that for a bit here.

Going back to images, evaluation for images is comparatively straightforward. Images typically-- We use images and inputs and then we use some sort of smallish number of classification outputs, even a large number, maybe even thousands of potential outputs for classification. But it's usually a discreet list. And this could be things like, does this image show evidence of a disease? Could be what level of severity of this disease does this image show? And the nice part about a small discrete number of outputs is there's pretty well worked out metrics and statistics for measuring performance based on those. We can take a bunch of images, we can look at the outputs, we can compare those to some sort of ground truth like expert raters, and we can measure things like sensitivity, specificity, basic accuracy.

The more complex versions are disease severity where maybe the difference between two options-- We sort of want to penalize more of the difference between high severity and low severity versus two grades in the middle. There are still pretty nice techniques for doing that, things like CAPA scores. But if we look at generative AI models, this is an example of an input and an output for that. The input here is a question, how do humans get toxoplasmosis? And the output is about a hundred-word description of what that might be. How do you tell whether this is right or wrong? Is this a correct answer or an incorrect answer? And to what degree is it correct or incorrect? You can sort of immediately see this is a very different problem to try and tackle.

There aren't well worked out methods here. There's no sort of simple widely accepted thing we can look at. And so, a bunch of groups have been looking at, okay, what are these? Can we build custom multidimensional frameworks for how we think about what makes an answer right or wrong? These are examples from a paper from ours from last year where we looked at a battery of clinician evaluation frameworks where we looked at things like extent of possible harm, evidence of incorrect reasoning, is there missing content? But these metrics were arrived at based on the intuition of people who are first looking at this problem, and I think there's a lot of work to be done here. Those are all looking at the answer from a clinician framework, but in general, these models will be used in products that will go to people and people might approach this and have different answers of what the right answer is. For example, not only does the answer have to look good to a clinician who's assessing the correctness from the expert's perspective, but it also has to actually satisfy the needs of the user that it's serving. Does this answer capture the user's intent? Is it answering the right question? Is

the answer framed in a way that's helpful to the user that helps them solve their problem? These kinds of things. You can see it gets very complex very quickly.

Taking those parts and kind of putting them into the larger picture of what does this mean? What are the big challenges going to be in evaluation here? As I mentioned before, the crux of the problem is kind of this difference in the number of outputs. For predictive AI, we have two to thousands of classifications. For generative AI really it's almost infinite. It's when you start thinking about longer and longer text as potential outputs. The number of possible outputs gets astronomical very quickly.

The metrics are fairly well worked out for predictive AI. We understand sensitivity, we understand specificity. These things that are familiar to us. generative AI, there's a large number of metrics that are being used. None of these are well accepted and widely accepted yet, even the rubrics on which are measuring things are not accepted yet. So, there's a lot of work to be done there. Similarly, for predictive AI, one of the things that is important for especially postmarket monitoring where you're thinking about deploying these systems to larger scales and you're looking to measure whether it's doing a good job or not and monitor it, automation starts to become important for that kind of scale. And predictive AI is very automatable. You can count the number of times you got it right or wrong and you can ship a thousand images over with their grades and quickly have a computer go calculate what percentage you got right and wrong. That's not so easy with generative AI. If you've got a whole bunch of text strings that are coming out of this, you need that human judgment element right now to go measure that, and that's slow, it's costly and it doesn't scale well.

But how do we expect this to go over time? And this is a little bit of sort of my personal prediction here. I think we'll see pretty rapid progress in the coming months and years towards a small number of metrics and standard valuations that do become stable and widely accepted. We figure out a way to make-- We probably figured out a way to make it cost effective to do this at some sort of scale, have some level of automation in there, but I think it's likely that those are really only going to happen in the next several years for a pretty small subset of use cases for generative AI that are the most broadly deployed. These are going to be areas that get the most funding and the most interest and the most research, and I think full robust evaluation of all the use cases we expect to see developing over time, that's going to take quite a bit longer. In particular, when you start to think about things like multimodal AI where you're bringing in pictures and images, and videos into this evaluation framework and kind of running into the same problems but with whole new forms of media.

I think it's a great place for us to be investing. It's very important. It'll be key to making sure these models are deployed and serve the world in a responsible way. So, I hope to see a lot of work there in the future. With that, thank you very much for your time. Take care.

Dr. Bhatt: Great, thank you so much. That concludes our speakers for this morning. I will do a quick summary and then we will have time for questions from the Committee for the speakers and some discussion. We are aiming for lunch at one o'clock just to give everybody a sense of how much time we have for discussion.

In the second session, we started with Ms. Paulsen from CDRH. It's encouraging to know that the FDA is working to globally align with other regulatory

efforts in AI, so thank you for leading with that. It's also good to know that there are not yet any GenAI or adaptive AI approval, so our timing is perfect and our group feels relevant today. We discussed controls for postmarket monitoring, such as the need for a performance monitoring plan, and we got a nice in-depth look into the predetermined change control plan. This includes plan changes to devices if the device would remain safe and effective obviating the need for a supplemental application. It is applicable to all devices right now, and it might be a nice way to allow the improvement of GenAI-enabled devices over time because we know that generative AI will get better at a rapid pace, but it does create requirements for safety and efficacy evidence meeting a certain level to pass those PCCP criteria. The PCCP includes a description of modifications, manual or automated, implemented globally or locally as we've been talking about sites a lot and requires that we maintain the device within its intended use.

I'll put to the Committee, is there a modification of this statement for generative AI that we should talk about later? The modification protocol includes verification, validation, data management, retraining, performance evaluation and updated procedures. A lot of the things that we've been asking for from generative AI. Further, could the FDA PCCP impact assessment that was mentioned become an interval necessary for protocol post-deployment for a pre-specified period of time? As was discussed by Steven earlier. So, I'm excited for the consideration of the PCCP as we get to our afternoon.

Dr. Longhurst gave us increased authority-- No, he suggested we might consider increasing authority to require transparent outcomes data and postmarketing surveillance. The current study and reports of generative AI for usability, successful implementation or clinical outcomes is lacking according to his presentation. And I

challenge the Committee to directly address whether those concepts might fall under FDA purview in the premarket and postmarket stage. The AI principles by Dr. Longhurst generative AI for example for in-basket responses. If you start with a draft, it transparently says you start with a draft. And then I was excited to hear about his network of implementation science centers rather than for-profit assurance centers. Interesting to think about embedding that along the way.

Dr. Kottler then gave us the specifics of GenAI in radiology and the need for specificity of what kinds of errors occur. The language for errors will help us educate users to mitigate risk. Her GenAI error examples, I'll be honest, gave me palpitations. I don't just say that because I'm a cardiologist, but I will look at the Committee and say, we really should think about whether education to use generative AI might be a need to have rather than a nice to have because that is complex and it was for us.

A framework for continuous monitoring, concordance or discordance with report results, and then some level of automated or clinician in-the-loop adjudication are great mechanisms we can talk about this afternoon for postmarket surveillance. We also learned about ensembling, which is an embedded quality assurance check, which I think is a great opportunity for us to talk about later. And lastly, look at the distribution of real data versus training data. We talked about labels saying what the data is trained on, but the idea that you may not actually know whether that's your data unless you look and see whether it's similar was really important.

Model stress testing user feedback came up and our colleague from Google further demonstrated mechanisms from model evaluation, how to evaluate qualitative answers with multidimensional frameworks that they're still working on. Does the

answer capture the intent and is the answer helpful? That almost felt like a net promoter score, but perhaps was one of the most helpful things on that slide. So again, something for us to think about.

And I want to end with our patient-centric approach to postmarket surveillance. Dr. Cordovano, thank you. An intentional partnership with diverse populations and patient voices across a total product life cycle, access to the AI output as part of the medical record in a transparent fashion that can inform shared decision making. Your phrase, "Ma'am, we don't do that." I'll take a moment. Dr. Cordovano, you shared something personal. I'll tell you-- I had completely forgotten. I had a report that said one thing and a doctor's letter that said another, and I now recall how excited they were when I went in to tell me they were using AI. I still probably need to go back and figure out why the differential occurred. It happens to all of us and it's our information. So, thank you for your presentation.

Lastly, I do like the idea of a patient task force and the patient as a human in-the-loop. I'll take us back to CTA. That was our first presentation. One of the first times Kerri Haresign and I were on a panel together. We talked about the fact that it's not consumerism, it's patient agency. I think this was a great session today and I'll open us up for any questions for the panelists. Dr. Elkin and then Dr. Shah, then Dr. Maddox.

Open Committee Discussion Q&A (*Clarification questions*)

Dr. Elkin: My question is for our Consumer Health Advocate speaker. We talked a lot about transparency in understanding whether AI is because of safety issues were involved in people's care or communications. But one thing you didn't talk about is the

demand that AI be used in their care if it shows that a clinician with AI is superior to a clinician without. Could you comment about that?

Dr. Kottler: Hi, Nina Kottler. I'm not sure if you're directing that at me, but I can provide an answer. Nope, someone else?

Dr. Elkin: Yeah. You, if it's okay.

Dr. Kottler: Okay.

Dr. Cordovano: Thank you for that question. I ran into this. It actually makes me think of second opinions or third opinions when patients living with cancer have progression or need additional advice and certainly are looking for something else and looking for facilities that may have a different technology or may have AI that they could apply when maybe they've been sent home to die. There's not even enough transparency, and I'm really good at what I do. I still can't find a publicly available resource that lists all of the different 950 technologies that are approved and who's offering and what provider's office, and what cancer centers. So, I can't even direct patients to that.

And the challenges become a lot of the second, third, and fourth opinions or traveling for clinical trials requires an investment of time, resources, childcare, days off from work, applying for disability. So, these are really, really massive decisions that people need to make and we're blindly making them, assuming a facility has a particular AI power tool, but we don't know, is it the same one that may be, for example, my local cancer center is using? Or is it the same one that Sloan is using? Do I want to go to another facility that has that same tool or do I want to go to a facility that has a different one to get an authentic second opinion?

So many questions. I wish I had a better answer, but I will say we need a publicly available resource to now map where are those 950 approved tools being used, what providers and provide transparency there. In an ideal world, what patients really want-- We talk about a nutrition label. I want to go to my hospital and my cancer center in the same way that there's bios on a website, create bios for all of these tools that have the nutrition label information, but also who's using it, what are the outcomes, how often, what is the volume, increase in cost, decrease in cost, a lot that you could do there to create more information for the public domain.

Dr. Elkin: We have been talking about model cards that would be available with respect to any model that's out there that should be available publicly.

Dr. Shah: My question is for Dr. Kottler. It's very specific. I really appreciated your presentation because it was a real-world study done with a large number of samples. When I review for the NIH, we have rigor and reproducibility and we also have the power of statistical analysis, and you also have specific examples of what was going wrong. So, statistically, and then we can go to the human element, you are seeing both type one and type two errors. Type one errors are false positives. Type two errors are false negatives, right? It's going in both directions, upper and lower bounds. My baseline common sense question is how many of those errors would clinicians make by themselves? Both type one and type two. What's your ground truth? What's your baseline specifically for those type one and type two errors you're seeing?

Dr. Kottler: Yeah, the literature suggests a very large variability in that number. Anywhere from five to 30%, you can see in literature. The data that we have showed at least the 1.7% error rate that was caused by the clinician because that was mirrored by

the AI and we were able to find those. Otherwise, you'd have to go through every case and do that. I think it's one of our next studies that we will do to exactly understand what the error rate is. The question is the error rate is going to be something that is very different location to location and person to person. What we're using right now to understand is a system that's not really working and that's peer review.

Dr. Shah: And my point is what you're saying that essentially a clinician could be actually performing much worse and you can have an augmented performance because of the algorithm, right? That's a possibility counterfactually. And you could have an inverse error where the algorithm is performing poorly and the clinician is helping it perform better, right? So, you're having those kinds of things. And so that question that leads to the very patient-centered question ahead, what are the outcomes you're measuring? Are you measuring efficacy in clinical decision making, reducing clinical time taken to sign off a report or are you tracking downstream if the report is generated? Are there extra tests ordered? What is the outcome?

Dr. Kottler: Yeah, great question about outcome because ultimately that's what's really important. I'm going to talk about your first point which is what are the differences that you're seeing? Sometimes you could put a human with an AI system and that could improve things. We have been doing this for seven, eight years now, and that is exactly what we've seen. If it's the right physician, an educated physician, that the expert plus the AI is better than either one alone, especially if you are training that expert over time. And we've seen that on convolutional neural networks. And I'll give you one example. There's a model that we have looked at that looks for brain aneurysms on a CTA of the head, and this is for any CTA of the head often done for stroke. And what we found is that the AI helps improve the sensitivity of the radiologist

by 24%. But if you look, you can also find that the radiologist would've helped the AI by 34%, and it's always a bi-directional thing and it's because the AI is looking at things in a different way than we are. And when you bring those two curves together, hopefully they don't overlap perfectly because it is when they don't overlap that you get that benefit. And if you're finding a perfect overlap, then there's no use in having the AI tool, just adding cost.

Dr. Shah: What are your outcomes?

Dr. Kottler: Outcomes are a harder thing for us in radiology to measure at scale unless we have access to EHR data. And that is something that in a hospital system that's not always given to us at scale. So, we look at it in a few different ways based on the information that we have. It doesn't mean that we shouldn't be evaluating. Now, value is in the eye of the beholder, you have to decide what beholder you're representing. As a radiology practice, right now, the biggest problem in any radiology practice is what a bunch of people mentioned, which is how efficient can we be? And how much mental burnout can we decrease? That is if the radiology practice is buying that tool. So, we absolutely can measure and do measure that. That's actually the minority of benefits from the AI that's out there today. I think it will be a much greater benefit with GenAI.

Dr. Shah: So, your outcome is mostly clinical facing, it's improving the efficiency of clinical decision-making.

Dr. Kottler: That's one output. A second output that we look at is if you are a hospital, the efficiency is not as big of a value for you, but where most hospitals care about-- Now, everyone cares about quality and we're measuring quality as well, and

you want to improve quality. But what a lot of hospitals are looking for is there a return on investment? No one has dollars to pay for this. How can it get out there? Part of why it's not getting out there is because there's no return on investment. So, we're looking very specifically depending on the type of hospitals at fee-for-service, is it value? Is there value being provided? I would love to be able to do more outcome types of metrics, which will take longer to get, right? These can give us information now if I can get diffuse access to the MRI data.

Dr. Maddox: If you don't mind staying at the podium. I just wanted to very much sort of think along the lines of Dr. Shah. I'm trying to get back to the access to the workforce stuff that we've talked about. In my opinion, a lot of these approaches are going to be incredibly helpful where we are short-staffed or are absent staffed of radiology interpretation. So, although I agree it's going to be important to understand person, machine, person plus machine, I think that's largely in the domain of these large centers where a lot of our work is occurring. And I wonder-- I'd just be interested in your opinion as we're thinking about advising about postmarketing surveillance. I look at your results, roughly 5% of the GenAI tool getting an error rate that drops to 1% when you have an expert in-the-loop. But the scenario I'm thinking about is that primary care person or ED physician, or provider out in the community who's going to largely rely autonomously on the interpretation of this tool and what error rate is acceptable. And so, I wonder if that's one way to think about it, at least for this part of our healthcare system and connected to that, do we need to think about further fine combing the way we measure this as not clinically significant and we could argue a little bit missing left or right.

I'm not debating its importance of including, but I guess I'm wondering about more of the outcome around-- Is our recommendation going to be to continue care for a finding or are we giving reassurance that we're good for now? That to me is really what the question is in front of our particularly under-resourced medical centers. You're right about EHR, but in some ways that ultimately seems to be the outcome from the radiology report that I, as a cardiologist in my example, tend to use. I'm like, "Oh, is that nodule? Do I need to do anything else or are we good?" And if we're good and we have low error rates there, that seems incredibly helpful for particularly places that are under-resourced. So, I wonder a little bit about your thoughts on that construct and ways that we might advise around those types of surveillance activities.

Dr. Kottler: Sure, yeah. I mean you absolutely should look at 5% and say, "Well, that's not that high. Maybe that's acceptable." But if you look at the type of errors where it's changing a stenosis from 40% to 30% or 30% to 80%, the outcome on that patient is significant and that's not acceptable. And that's something that we can figure out at the point of care. It's not that hard for an expert to do it. Now, we have to solve the problem where we don't have enough physicians or technologists to go around. And that is a problem that if we don't solve, you're right none of this is going to be useful. We're going to end up with care that is maybe a little bit better like an emergency department or a nurse practitioner. They're going to do better with AI than without it. But is that good enough?

I mean, if we're trying to improve the quality of care we're giving today, we should improve it from the baseline and the baseline is an expert read. So, we have to solve the problem of this disparity between volume and capacity. And while none of the AI thus far has really done it, it doesn't mean that it can't. And there are absolutely

ways that it can be done. Voice recognition, which takes about 50% of the time that we spend as radiologists, is done in the voice recognition component, the dictation component, and the administrative tasks. If you take away that part by automating components of that, you're significantly going to improve the efficiency of the radiologist. We absolutely can do this. We will be doing it. You're going to start seeing some this year and a lot more next year. And if you say we're maybe 10, 20% understaffed, if you could improve the efficiency of all of the radiologists or many of them by even more than that, we're going to have enough staff. And I'm actually not worried about the amount of staffing we're going to have in the future and that will allow us to maintain or improve the quality of care, but also be able to be effective in multiple different locations.

Now, the other thing that we haven't talked about is yes, the radiologist is an expert, but we also have subspecialist expertise within radiology. Radiologists that are trained specifically in cardiac imaging or specifically on neuroimaging. And those radiologists are even better than general radiologists overall. And what we found is that the AI helps the general radiologist get to the level of the expert radiologist. That's not necessarily the case with a general physician.

I think we're going to solve this problem. We are in a time period where right now it's a problem and a crunch, but it's going to get solved I think over the next year or two. And I think we should always think about improving the quality of care rather than changing what we're doing to mitigate an issue that I think is going to be resolved in other ways. The other thing you mentioned, what was your second question? Outcomes? No.

Dr. Maddox: Well, I think it was-- Sort of thinking about outcomes a little bit on the suggestions for further management or not as opposed to some of the smaller details-

Not smaller. The other details you mentioned in your evaluation,

Dr. Kottler: Right. A lot of what a radiologist does is try to make a diagnosis or a differential diagnosis and then suggest based on that patient the exam and the diagnosis, that pathology, what should happen next. And that's why we're the clinician because the goal is to help you all as referring clinicians make a more informed decision. And if the data that we're using or that is being used by the AI to make that decision is wrong, or if the recommendation is wrong, you guys are going to get the wrong output. And how often are you checking to make sure that the recommendation is right?

I mean, even for me at the point of care when I'm doing this, there are population health guidelines that are out there in journal articles. And in order to implement one, number one as a human, I have to remember there is one. Humans are terrible at remembering. Number two, I then have to leave my workflow, look up that recommendation, follow the crazy flow chart, make sure I pick the right one and then dictate that. Right now, we have clinical decision support tools that are helping us with that because we're not that great at it. My question is if a wrong result goes out because of AI, is the referring clinician going to go through that same process? I'll tell you it doesn't happen in radiology.

Dr. Maddox: Thank you.

Dr. Bhatt: Thank you. We're going to go to Dr. Soni next, but somehow sitting next to Dr. Maddox and his contrarian views yesterday it's rubbing off. So, I will offer maybe a few things. One is we talk a lot-- I'm feeling the digital divide and so I'm

saying that out loud, right? We talk a lot about our ability to have the exact right data go in and the expert look at the data come out. And that doesn't happen in a majority of our country. I think the second part that we as the Committee need to think about therefore is this question of triage, sick or not sick, and does that deserve maybe a different discussion this afternoon than the exact quality of the report? And what does that look like? I don't have an answer, I'm just putting it out there so we remember to talk about it in that version this afternoon.

And then lastly, I will say when we talk about the result that we're getting, being compared to an expert read, there are many places in this country not getting an expert read. And so, the result we might be able to offer using some of these augmented devices with generative AI may be much better than the care that was being received. We have to be very careful that we're not taking the risk there rather than the benefit. I don't know how to do that, but I want us to talk specifically about some of those underserved areas this afternoon as well because it might be a slightly different discussion. I had Dr. Soni. Dr. Soni, good timing? Okay.

Dr. Soni: Sorry, I had to step away to take a call. My question is for Dr. Paulsen and its surrounding shared responsibility. Some of these interfaces with the comments that we had at the closing of the last session, which was really helpful to think about and I have been gaining a great appreciation from a lot of the radiologist colleagues about how different radiology and GenAI are, and the rest of the medical specialty in GenAI is. And what I've been reflecting on are my peers in primary care who are spending more and more time doing administrative and billing tasks, mostly for legal and billing reasons, and are spending a copious amount of time outside of their clinical practice.

Job desk to pajama time to the point where children are wondering why their mothers are not spending dinner time with them because four to five hours go into pajama time.

Chris mentioned some of the advances in the use of generative AI to respond to patient messages, to respond to clinical notes. The challenge becomes what responsibility electronic medical record companies have as a shared responsibility for enabling generative AI. Especially when we think about hospitals that don't have Epic and Cerner, the two largest market shares of electronic medical records. Are they going to become less advantageous for being able to deploy some of these technologies? Conversely, if Epic and Cerner, and some of the other companies have preferences or their own products, sometimes the market change or the pre-planned change, are they going to be able to have any influence on how those changes get implemented? And what responsibilities do you envision electronic health record companies having?

Dr. Paulsen: Yeah, that's a big question. I might actually look to Sonja to share some thoughts on, again, how we view those companies and electronic health records in the space. Again, we're focused on medical devices, so I don't know if she wants to clarify our purview.

Dr. Fulmar: Yeah, that's exactly right. I think that while we're very interested in learning about the shared responsibility that everybody has in this space, including the health records system manufacturers, the hospitals, the care providers, and the device manufacturers are the ones that are within our purview. And so, the questions that we're asking today about what types of controls we might want to apply are going to be for those device manufacturers. Not to say that I don't want to encourage that conversation beyond that as well because I do think that there's quite a bit to do there. And I think

that when we're considering tools like the predetermined change control plan and other mechanisms that we have to require manufacturers to do something, I think it will require them, and I use the word "require" there, not in the legal sense. It will encourage them maybe then to work more collaboratively with healthcare systems, with electronic patient record systems so that they can gather the information that they need to fulfill requirements that might be in a predetermined change control plan or might be in a special control for collecting postmarket monitoring data.

Dr. Bhatt: Thank you so much. We'll do Dr. Kukafka, Dr. Elkin, and then lunch.

Dr. Kukafka: Quick. Okay. So, my question is for Dr. Cordovano. First of all, I really appreciated your talk.

Dr. Cordovano: Thank you.

Dr. Kukafka: And the patient perspective.

Dr. Cordovano: Thank you.

Dr. Kukafka: Because I don't think we focus that much on it, but I think it's critical. You talked about training and we did have a conversation about training, and the feasibility. I think of one of my colleagues, Jim Cimino, for the past 15 years, he's been saying patients don't go to patient school. So, are you thinking of patient training for this specific device or could you just talk a little bit more about what your vision is for training?

Dr. Cordovano: I can give you some examples from my patient advocacy work. Once ChatGPT hit, I got up to speed as quickly as I could and then started coaching patients on smart, safe prompting. Where do you start? There's still really not a white

paper or a resource available, but there are very tangible ways of going in, prompting, giving it a voice, giving it tasks. How do you pull information from a medical record? How do you bring it down to a fifth-grade level? To an eighth-grade reading level? How do you translate it into Portuguese? How do you then take it and compare it against Google Translate to make sure that it's close if you don't have access to a translator? So, there are certainly steps with training, and this is all informal, but could there be training that's developed? We're hearing about the necessity of training for physicians. We certainly don't have enough resources for patients.

I know a number of nonprofits and patient-focused communities are sharing best practices on how to use these different LLM tools that are currently available. Certainly, can apply to these devices and I would love to see device manufacturers and even pharma come in with their interpretations of how to help patients through this next era of using these tools responsibly, not just to use responsibly, but also contributing to the ecosystem with their findings and insights, and correcting errors and reducing harm and bias and risks. I think we're all quite capable of doing that as well.

Dr. Kukafka: Just to follow up, so do you think that that patient role would improve the safety of these devices?

Dr. Cordovano: Absolutely. Including not just the patient, but also families, especially when you have multiple comorbidities, something life-altering, life-limiting, that care partner, caregiver is often so helpless and wants to do something that they're-- Think about the person-- I'm sure many of you have been there. You may have a loved one in a hospital bed and you're just sitting there, and you're helpless and you don't want to leave because maybe you don't want to miss the doctor round, and what a great

opportunity to engage that person that's sitting there. And if you walk up and down the hallways, most of the rooms do have someone sitting there and we're not engaging them at all.

Dr. Kukafka: I think of one and then I'll keep quiet because we want to do lunch, but we know that patients could devote as much as 80 or a hundred hours a week, if not more, to become experts in their own care, which is more than a specialist would devote, right? So, I would agree with you that there's a strong role for the patient in ensuring the safety of these devices that we should pay attention to.

Dr. Cordovano: I love that you said that someone used the word expert in the loop, and we certainly need to include the patient's family as experts in their care to contribute to what we're trying to build here to care. Thank you.

Dr. Elkin: So, I just wanted to build on something Dr. Maddox and Bhatt said earlier that I think there's a very good clinical example that we use already and could be low-hanging fruit for what we want to think about regulatorily in this case. So, if you think about the use of highly sensitive D-dimers in VTE, we use them because they have a very high sensitivity or specificity for ruling out VTE. And then if they're positive, we don't know what the case is, but it can move on to something else. You brought up whether you have no-go decisions in what you need to do next, and there may be very high reliability in defining normal and getting people to realize that this case is okay and you really don't need to do 23 additional tests to know that the patient will be okay, because we have enough information from this screen. And I think that kind of thinking in terms of-- We're always thinking about making the diagnosis, but it's also very helpful if serious diagnoses can be ruled out in a particular case and could

lead to efficiencies in the practice cost saving for the country. And it could be a safe way to proceed in many cases.

Dr. Bhatt: I see the look in your eye.

Dr. Clarkson: Yeah. I have briefly a couple of thoughts I want to pull together in echo following from that. Patients want to know why decisions were made and this is AI independent, whether with or without. That's one big idea. So, if AI is overridden, it needs to be documented and like some justification that needs to be part of the record.

I also want to talk about rural areas very briefly. I am from rural Kansas. And to help you understand how these ideas come together, I'll very briefly tell you why I'm sitting in the patient advocate chair here. 12 years ago, my father, who lived in rural Kansas, was very badly burned. My mother drove him to the local, very tiny hospital. It was about 6:30 in the evening. Now, not only did this hospital not have an ER physician because it's that small, there was no physician there. They had to call someone in. So, the reality is that an enormous number of people in this country live in very, very rural areas where there's not even a physician within an hour's drive.

So, the nurses knew he needed to be transferred to a burn center, but the physician that was called in refused to do a transfer. Now, days later, I looked up the American Burn Association guidelines. It was very clear to me he qualified for multiple and he only needed one to be sent to a burn center, but he was not transferred. He spent the night in that tiny rural hospital. He received the wrong medications, the wrong fluid, he was conscious for most of this and in incredible pain. And again, the nurses knew he needed to be transferred. The doctor did not do it. The next morning, a physician's assistant walked in, realized immediately he needed to be transferred. At that point, he

was sent to Wichita about an hour and 15 minutes away. But at that point, damage was done and he, although undergoing a skin graft surgery in his entire back and a few other areas, did die about a week later.

So, this brings together the: why wasn't the transfer done? My family went through three years-- It took three years to go through a lawsuit and a settlement. We still don't know why that transfer wasn't done. And so, patients want to know why decisions were made in any context. That's one thing.

But also, this brings in the reality of rural America that there's a huge role for just: does this patient need to be sent elsewhere quickly or not? And I think the AI doesn't have to be perfect to have a role and that this patient can't be handled at this facility. It needs to go quickly somewhere else. Thanks.

Dr. Bhatt: Thank you so much for sharing such a personal story. And I think it gives us both great pause, but also a cause for when we come back after lunch to try and figure out how generative AI may actually help us enable access for those who need it the most. So, thank you. We'll now adjourn for lunch. I will ask the Committee to please not speak about this topic with one another or anyone else. And we'll come back in a half hour if that's okay with everyone today so that we can keep moving. Thank you.

Committee Discussion of the FDA's Questions (*Deliberation and Response to FDA*)

Dr. Bhatt: Okay, welcome back. We'll get started for our last session of the two-day meeting. It is now time for Panel deliberations and discussion of the FDA questions. The FDA has generated a series of questions for the Panel to consider. We are on number three today. It has three parts. Ms. Aubrey Shick will present the FDA

questions and we the Panel will deliberate amongst ourselves. I make a request, we're going to be together for a couple of years doing this, so you'll see I'll try and change things each day to get a little better.

For today, let's look at the main question in all three sub-bullets in the beginning so we know what they are. And then we'll start with A, then B, then C, and really try to orient our comments towards getting concrete answers to each section before we move on to another one that also makes it just a little easier on me in terms of transcribing our thoughts for the sake of the FDA if that's okay. Again, please say your name for the sake of transcription before your comment. And when you give your comment, please end with your concrete final sentence "In response to 3A, this is what I would like us to record for the FDA," and let's see how that goes. Ms. Shick, please. Thank you.

Ms. Shick: Thank you, Dr. Bhatt. And just to clarify, are you asking for me to read all three upfront? Okay, we'll do that. So, our final question area is in regard to postmarket performance monitoring. As we've been hearing about today, postmarket performance monitoring and evaluation may be important for these devices, particularly because they're non-deterministic. Additionally, after deployment, many generative AI devices will undergo continuous adjustment based on localized live data, user interactions, and changing conditions. Please discuss the aspects of postmarket monitoring and evaluation that will be critical to maintaining the safety and effectiveness of these devices.

So, for question 3A, we have: What specific monitoring capabilities should be considered to effectively evaluate and monitor the postmarket performance of

generative AI-enabled devices to ensure they maintain adequate accuracy, relevance, reliability, especially when adapting to new data?

Question B: What specific strategies-- And is that on the screen? Yes. What specific strategies and tools can be implemented to monitor and manage the performance and accuracy of generative AI-enabled devices implemented across multiple sites, ensuring consistency, and addressing potential regional biases and data variations compared to the device that was authorized?

And then C, the final part of the question: What methods and metrics can be utilized to effectively monitor and evaluate the postmarket performance of generative AI-enabled devices that use multi-layer application design, for example, the device queries external consumer-grade AI services that are not themselves medical devices?

Dr. Bhatt: Great, thank you very much. If we could put part A back up on the screen for the panel to reference. For A, we're talking about specific monitoring capabilities that we want to consider and we're looking at that with accuracy, relevance, and reliability, especially when we add new data. I'll open it up for conversation. Again, please say your name when you start for the transcript, and please end with "For part 3A, I would recommend the following."

Dr. Radman: So, let's see. Sorry, I'm just getting my notes together here.

Dr. Bhatt: This is Thomas Radman who'll be speaking.

Dr. Radman: Thank you. I really liked the phrase that the team from Deloitte used today, and it was "significant human oversight." I think that really encapsulates a lot of the challenge here. And then this is a reiteration from yesterday, but I think it may be 3A or 3B about having some regulatory requirement. And actually, the last speaker, Dr.

Cordovano mentioned her frustration that patients don't know what to do if they find something wrong. Someone yesterday mentioned the ability-- These devices are digital devices with Internet access, so it doesn't seem like a huge engineering lift to have some kind of mechanism within the device where it's at the push of a button almost errors or hallucinations are sent to a centralized independent database, and also the manufacturer just on a reporting basis. Like analogous to the FDA's adverse event reporting in a way, but just facilitated by the advantage of it being a digital device.

And I just wanted to mention that there is a large-scale citizen science project that is collecting an error database for ChatGPT. There's some information on it in an article by Harrer et al. And then also in the pre-meeting materials, there was the public comment docket and one of them was by this company Woebot, W, O, E, B, O, T. And they talk in their comment, in their docket, about internal periodic review of chat transcripts. Perhaps part of the FDA's evaluation could include the mechanisms a company has in place to do such a periodic internal review. I found the three really astute. Thank you.

Dr. Bhatt: Great. Just to report back, I'm hearing that the specific monitoring capabilities might include an internal periodic review of chat or interaction transcripts, error reporting back to the manufacturer, and perhaps a central database depending on the level of error. And then indications as to where significant human oversight might be necessary during periodic monitoring. Does that sound--?

Dr. Radman: And also, this idea of an automation or a facilitation of error reporting, especially when you're talking about a patient-facing device where it's like pushing a button. In fact, sorry, something I wanted to mention is that there could be a

requirement just like we do labeling studies in a premarket space and usability, there could be some demonstration of, whoever the user is, their ability to find the error reporting. We don't want it to be another one of these ineffectual labeling kind of things where it's like, "Yeah, we did the labeling," but in practice, people don't recognize it and we don't want this reporting system to be that kind of thing where it's in place, but no one can find it. No one knows about it.

Dr. Bhatt: Excellent. Thank you so much. Dr. Botsis, then Dr. Soni, and then Dr. Jackson.

Dr. Botsis: Thank you. Taxiarchis Botsis. So, it seems like there will be a need to capture additional information for GenAI technologies, and inspired by Dr. Kukafka's comment yesterday about health leaders in numeracy, I think we'll probably need to capture elements of that kind. I guess the challenge there, especially when we want to link that to existing data standards that are used in institutions, for example, observational data standards like the V-map data model. We have to examine whether those data standards can support capturing that, those additional data elements. And there may be a challenge there. But I think we'll have to start from there and really evaluate what additional data elements are really needed for that part.

Dr. Bhatt: I'm just going to repeat back, we already have data standards that exist for a lot of healthcare and therefore maybe to start from there and think about what additional data standards need to be included, specific to generative AI being in these devices. Is that correct?

Dr. Botsis: Yeah, exactly.

Dr. Soni: Just building on that, I think-- I continue to be really just impressed and in awe of how much we can learn from radiology in this space. I think Dr. Dreyer's comment today about how many years of experience radiology has had here is really informative when we think about what kind of monitoring capabilities can allow us to evaluate and monitor postmarket performance. There was a great presentation yesterday by ACR about the network that they're creating and part of what's enabling that are those common data standards and data terminology, and ontology. I know Peter mentioned things to that effect. As we think about other fields and specialties where generative AI is going to be applied, we need to come to an agreement on what those terms are and what those common data models are that can allow us to follow the digital data crumbs that all of these technologies are going to create so we can learn on it and be intentional about what that platform is, how can that platform be accessible to all healthcare organizations that are deploying it, not just the academic medical centers and what's the feedback loops from the end user, whether it be providers or patients or advocates that are able to provide the feedback back to those capabilities that exist of continuing to monitor the real-world deployment of these tools.

Dr. Bhatt: To confirm, I'm hearing that we like the use of the existing data standards, but it is important to recognize the common data terms that need to be used across the subspecialties and where GenAI is used. And then importantly, if we are going to engage in monitoring at a larger scale, it needs to be a low-cost accessible mechanism for all people to be able to both give data in and get data out. Okay. Next, we have Dr. Jackson.

Dr. Jackson: Thank you. Jessica Jackson. I have two overall comments and then two specific monitoring capabilities for 3A. The first comment is, I know this is probably

going to be a much larger endeavor, but I think it was Dr. Dreyer. One of the things that he continuously said that stuck with me is how we start to also rethink about reclassifying. I think one of the things for me-- And I've said this before in other spaces, but I think what we're seeing with GenAI is similar to an industrial revolution and we might need to think about a different language, right? I think about things like social media 15, or 20 years ago that was not a term we were using and now we have different ways to measure it that would not fit into any other way. So, I think especially as GenAI continues to grow, how might we start to think about reclassifying it, which might change postmarket performance? Because one of the things that just struck me as listening is it almost seems like we're trying to force it to fit into what we already have and that may not work out well for anyone involved.

The other comment I want to think about is I think there's been a lot of focus as we think about images, especially in radiology, I recognize that they have the most ML AI-enabled devices currently, but I do think that GenAI will help other specialties to catch up and I want to be cautious of over skewing into thinking about what works specifically for that specialty and also thinking about what will enable other specialties to be able to leverage AI instead of being boxed in by some of the other things that are specific to something like radiology.

The two specific monitoring capabilities that I think should be considered, one is monitoring data drift. The capability to monitor data drift. When you think about the information changing for inputs, I think about something like digital therapeutics, which are authorized by the FDA and there's a lot of work in psychiatry and psychology looking at biomarkers. So, the data that we put in now for potential software as a medical device that could be GenAI-enabled will change as we get more and more into

the biomarkers compared to the data we're putting in now. And I think if we are not monitoring what data drift could look like, that will impact how those devices are used.

The other performance-- The other monitoring capability I think that we should have is the capability to consider the difference in product failure versus implementation failure. So, when we're looking at postmarket analysis, are we looking at: is this a failure of the product because the product itself is having issues or wasn't programmed right? Or is this an issue of the people that were not trained to use it correctly, which is an implementation failure when we're looking at what the analytics will say once something is out there? And I think there will be a big difference in that because we talked a lot about it yesterday about the training and the importance of it that will impact the performance of a product if people are not correctly trained. Thank you.

Dr. Bhatt: Fantastic. Thank you for your thoughts about being able to apply this throughout the practice of medicine. I think that's really important for us to record. And then specifically for our list here, I have monitoring data drift, specifically being ready for increased multimodal inputs as we're talking about biomarkers and other things. And then I love this idea of product failure versus implementation failure because there's one thing where the technology doesn't work and it's another when we can't use it. It goes back to the question of usability yesterday. Thank you. Dr. Elkin.

Dr. Elkin: I think that if we set up the right monitoring scheme, that'll help to address Dr. Jackson's concern about implementation versus the manufacturer of the activity or the artifact. And what I mean by that is that if we put in place the right data elements that are reported-- And I'd like personally to see them randomly reported for a percentage of the use as well as when someone thinks there's a failure because we all

know that when you have voluntary reporting like for vaccines and so forth, it only is 8% of the actual true cases that get reported, and that's really not enough to really get the full picture. So, I'd love-- And I think Dr. Radman said this, that because these machines can actually, they're on the Internet, they're in the cloud and they can do reporting so that we don't need to burden the user. We can put the burden on the manufacturer as part of their activities of daily living with the device.

And I think that if we came up with the right design for all of the data elements, we think are important, it would include all these features that we've been talking about that have characteristics like the usability of the system, who's the intended user, what's the use case that's being solved? Is this within the use case that came before the FDA or a different use case that came up in the particular situation that's being reported? And then enough identifying information that if we have a national repository of clinical information, we could get not just process measures but we could get outcomes. That is really where we all really want to be. We want to know that these devices are safe and are creating good strong clinical outcomes for our patients and their families, right?

And so, we need to think-- It's on us, I think, to come up with that and to make it happen. And then once we give you good recommendations, then Mr.. Tazbaz has to go and implement the thing in regulation, right? But in any case, I think that as a group we've delineated quite a number of the really essential foundational concepts that we need in order to do postmarketing surveillance, and it should inform our premarket approval over time right, because as we learn more about these devices through our surveillance, it's going to understand where we have insufficient premarket approval methods. Thank you.

Dr. Bhatt: May I ask you for a little specificity on that? Because I love this concept. When you gave us the idea in the premarket conversation about the model card, you were very specific about that. Right now, thinking about postmarketing surveillance to you out what you've heard the group say, if there were three key things where you'd say these are the three where we would start, may I ask you? And if not, now I can give you time and come back to you.

Dr. Elkin: No, I can help with that. There's real precedent. So, each of the large language models has a model card that's associated with it now, and that will have to be expanded for clinical use. But there should also be a model card from the data that's presented for premarket approval that comes out so that anybody who wants to use that device can see the kinds of details that we would need for postmarket approval. If you want me to give you specific data elements I can.

Dr. Bhatt: I'm going to ask on the postmarketing side for today. You gave me a very good list yesterday for the premarket, but for the postmarket side, in your mind the first three needs that would come up, what do you think those would be?

Dr. Elkin: Accuracy, safety, and then the bounded use case. So, if people are using these, and they will use the devices off-label, we've already established that in an earlier conversation. We need to know about it because if we're assessing failures, we have to know the scope of those failures. And this is how black-box warnings and things come up in postmarket surveillance, right? People identify a problem that was not seen in premarket approval that is an important piece of information for any practitioners who are using this. And then of course, if these are being used by the public, we have even higher standards of transparency in terms of how we inform people so that they can

understand it. People talked about the eighth-grade reading level. The New York Times is sixth-grade level, okay? USA today is third-grade level. So, I think we need to keep this down around the third-grade level if we're really going to talk to everybody in the country about the risks of these artifacts.

Dr. Bhatt: Wonderful. Thank you. We'll go to Dr. Radman next, and I will just say if anybody else has must-haves that are not accuracy, safety, or bounded use, please feel free to add that to your answer. Dr. Radman.

Dr. Radman: I wanted to bring up the topic of watermarking the outputs of these devices because there's a concern of poisoning the well, if you will, where we have devices that are influencing and inputting data to the electronic health record or other data sources, generating reports and so forth. And then, in turn, we're talking about continuously learning devices that are going to be using those same data sources to learn. And the effect that has is a convergence into a homogenous data source, the EHR because the outputs are biased. So, over time, you're just going to have a really monoculture, if you will, of the EHR where-- And that's going to further the digital divide and the disparities in health. This watermarking allows an easily identifiable data source where, "Okay, that was generated by GenAI. So, don't include that in the training data." And I think that's a really important requirement because we're just going to corrupt our electronic health records. It's going to cause immense patient harm.

Dr. Bhatt: Ms. Miller.

Ms. Miller: Thank you. I know it's hard to look in this corner, but I've been patient for a while. I have a few thoughts that I gathered and I'm going to start saying them all, but I just can't stop noticing that we are in this position of boiling the ocean at this

Committee and we have to do it. I mean, somebody has to start somewhere, but I just want maybe as Dr. Taxiarchis earlier today just focus on what is doable and what's not adding a lot of burden for all of us. So, I want to start by saying-- Appreciating the comments that we had today for PCCP and it's very good to hear and educate all of us about what is actually existing and what's in work in regard to the regulatory framework, and what we are starting to put together and working towards.

What I'm getting at is we don't want to add a lot of conflicting regulatory requirements that will create more confusion. And I think we should start by looking at what exactly is out there for postmarket monitoring. And you probably know more than I do, but especially for Class III devices, there is postmarket regulatory monitoring guidance that we have. And a lot of times there is conditional approval, premarket approval for a Class III device conditioned by a postmarket monitoring plan. So, that already exists for Class III high-risk devices. And we should start looking at what we have. Also, the quality management system that we as manufacturing implement today has a lot of these controls, special controls and post-marketing strategies in place. And I know, if I recall correctly, somewhere in 2026, the FDA intends to implement the quality management system regulations, which is even closing more loop with risk and monitoring and all of that.

Also, we have a medical advisory call and we have the CAPA process corrective action, preventive action process that ties the loop in postmarket monitoring back to how we correct our design controls and development practices because it's probably obvious, but I'm going to say it anyway, postmarket strategy starts with design, okay? The controls you put in and the rigor you put in your new design and incremental

reviews and all of that along the way, this is what builds up to that postmarket monitoring strategy.

Now, as I saw in PCCP and we know that because we review it, I mean in industry, we reviewed it before and we provided comments to PCCP separately, but there is a very clear section on exactly how you would do the changes and what you'd allow, and the impact assessment of what the changes will look like. And the range where you are allowed to implement those changes is a similar framework. And that framework can be expanded. And you said it very well to this kind of concept to where I know there's output variability, we can bound that by a concept like PCCP and go from there.

And I also call out the comment that was made today by our Google. I'm trying to remember his name, I'm so sorry, colleague. He very clearly articulated that there are very few use cases of GenAI today in medical devices. And those included in medical devices are none as we know, they didn't get approved. I mean we don't have any and want to point out that it'll take-- Can you hear me still? It will take a while until clear use cases where the manufacturer will be sure that there is no risk, no harm involved and the device is safe will be developed and metrics will evolve and some metrics that we look at today and we saw even hallucinations, how many kinds of hallucinations can be, and it's very important to be domain specific. There will be a lot of metrics that will be domain-specific by the time that the device makes it and then implemented in that for that particular use case and in that particular environment. So, we should make sure we leave room for specific domain metrics there instead of imposing, "Oh, this is how you should go about it," because it might not make sense for that use case in the future.

My final point, and it brings to the digital divide and I started by boiling the ocean. We have a divided ocean. Jesus' stuff. And there are comments today where how good is good enough and what low-risk applications like class I, and class II, can we just start getting our feet wet where it truly makes an impact in rural areas and does the benefit of bringing this technology to those they have nothing today. Having something with some error not perfect because humans are not perfect either might be beneficial and maybe this is a lower bar to where we can get learning and get more data from them. And we in the industry can then make a judgment call that "Yeah, we have enough data to do something bigger and better." And I think it is not random that radiology gets data collected for the past seven years and now we're able to look at the metrics and draw some conclusions and learn from it. It'll take years to make a really high-risk GenAI device. Thank you.

Dr. Bhatt: Thank you so much. Can I follow up with just one or two questions? The first is for the PCCP. We recognize that generative AI itself may change. That's kind of one thing. The second is the indication for which a company may create their model, may expand. Would you just give me kind of perspective as the industry voice here on which one of those two you think might be, or both, a reasonable place for us to start using PCCP as a base to say in the future if you want to change? And then, if anyone else has input.

Ms. Miller: The first step I think where we stay within indication of use and model changes-- And I think that's a current approach. I know we talked in the future to include the next step where we can expand the indication of use. And I asked yesterday for clarification when somebody was talking about expanding within that intended use to a different domain by same use-- Different-- Same use but different domain. So, that

nuance is there. I mean, it depends on how you look at it. Is that a new use or is the same use but different domain? So, the intended use, you can look at it in millions of ways. So, just-- That's what I'm going to say now.

Dr. Khubchandani: Jagdish Khubchandani. Diana, I agree with you. The metrics may change, but when you do a checklist before submitting an application, you mentioned my device, "Validity is this," "Reliability is this". And in the monitoring, we have to see how true it holds in the real world. The reliability proposed in validation studies, how reliable is it now? And those kinds of things. So, those could be standard. And then we add extra criteria for how this device now in the real world improves the quality of care or reduces the cost of care, or improves access to care. But I would hope some-- Like a risk-benefit ratio proposed in the clinical validation studies, does it hold true in the real world and does it hold true for all types of patients?

Ms. Miller: And it depends on the risk and-- Depends on the risk of the device. Sometimes we need to do those postmarket studies. I mean, this is a requirement today and it's not only for AI-enabled devices. So, it holds true even today for a high-risk device, and we have our part of the quality management system, we have signal triggers and triggers and reporting and medical device reporting and all of that. So, you do watch how the postmarket behaves. But I want to be very careful because this will bring-- There are manufacturers out there, so it'll bring a different disparity of what exactly the infrastructure that one can use is. And it's true that you need to afford the infrastructure to monitor, but if we want to go too much ahead of ourselves, then people are not going to make these devices because you can't develop that strategy. There is an infrastructure work that's required, I think, to implement a lot of the things that we talked about today. And yeah, there is risk benefit.

Dr. Maddox: I want to build on your comments about the change control plan. I wonder-- And I don't have a ton of experience about change control, so I certainly rely on our experts in FDA and elsewhere about how to think about this. But it occurs to me there are slightly different concepts between a product change control plan and the usage change control. And it may make sense to separate those concepts as we're designing a postmarketing strategy on the product side.

And again, all this should be sort of calibrated by risk. It feels like it would be worthwhile to think about "Can we specify a range of acceptable changes for the training data, input variability, and output variability?" And that we could decide you would allow for a greater range of those variables in a lower-risk use of generative AI. And then, if you exceeded those bounds, then you would need to explain, undergo the change control plan and the subsequent validation that might need to occur on the usage side.

And again, I can't speak with a great deal of sophistication on how the FDA thinks about deviations from on-label versus off-label and those usage changes, and when it makes sense to say "This is simply an expansion" or "It's appropriate for 510(k) pathway" or whatnot. But it does seem worthwhile to have a mechanism to at least in broad strokes, track the usage and the inevitable evolution of that so that we can have some idea of how much of a change-- If we're just going to a different domain, maybe that's less consequential than a totally different usage.

So, these are all variables that occurred to me. And I think the specific monitoring capabilities may be elements around change control plans for product and

then for usage and how we allow those to vary based on the risk. It's a recommendation I would add to our list. Thank you.

Dr. Soni: Apurv Soni from UMass. I really want to hone in on the monitoring capabilities surrounding the relevance part of the question. And this is where I want us to really, as a Committee maybe and myself, make a comment and an impassioned plea about really thinking about where those monitoring capabilities exist. Because in very simple terms, if we think about narrow AI or predictive AI as analyzing existing content, generative AI is very different where it's generating content and analyzing the existing content. And then the natural question then becomes "Where does that generated content recite and how is it relevant? How is it changing? What decisions are being made?" So, I really appreciate Dr. Radman's comment about watermarking that generated content so we can continue to evaluate it for accuracy, relevance and reliability. But the relevance part is really important there because we can see, and it's not impossible to imagine a future where a lot of the medical record is bloated by content generated by this generative AI and how much of learning, future learning, future reliance of models is off of this generated content.

There have been many dystopian but true stories of administrative tasks, like prior authorizations or requesting different components, that go back and forth between some of these generative-AI tools. And so, we need to be careful about what kind of oversight exists in the natural environment in which the content generated by generative AI is residing and how we can have proper provenance of that.

Dr. Rariy: I wanted to add two things. The first is more around how we're classifying the postmarketing strategy, or the postmarketing performance plan, in that I

do think we have a number of speakers today, Dr. Dryer being one in particular, who highlighted the importance of perhaps expanding the way we are describing postmarket performance. Meaning if there is a concept around pre-deployment, what does that monitoring look like? Perhaps more upstream in addition to continuous. And so, I think we should consider looking at what the continuum of monitoring looks like and that would include the pre-deployment phase as well. And one suggestion was the clinical validation, but there are a number of other strategies that we've already spoken about about what type of monitoring capabilities would be included at these different stages.

So, that's one comment that I want to make sure that we capture. It's not currently captured, but there's a necessity to consider that continuum.

The other, more specific to A, when we're looking at different strategies or different capabilities that need to be included, I think it would be important to create or facilitate an opportunity to have random auditing, if that makes sense. So, random external auditing of the device or the product if a certain threshold is hit. But I say that in the sense of-- Or even not to get the sense of a big brother monitoring in the background, but we do have to have some opportunity, I think, to have deeper insight into the device regardless of what information we're given. And I think that could be done both on a local level, more decentralized perhaps in addition to a more national level if we wanted to create some autonomy around that monitoring or that auditing.

The other, I would recommend specifically looking at various capabilities around bias and equity. So, specifically calling out the necessity of the device to monitor this. And I am assuming depending on the device, depending on the use case, there will be different mechanisms on how to monitor this. And I'm sure they will

change and evolve over time. And so, I'm not suggesting that we, the FDA, perhaps dictate exactly how to monitor equity and bias for that specific device, but that we certainly recommend that such a plan exists.

And then lastly, I know we have discussed the AI-assurance standards in general. I just want to make sure that we are cognizant that depending on how we classify whether there're five or seven key components of an AI-assurance standard, but that there is a classification and a framework by which this postmarketing monitoring is held is going to be important as well.

Dr. Bhatt: Okay. Ms. Shick, would you mind showing us Number 3, Part B, please, and reading that? And then we will take the comments regarding B.

Ms. Shick: This is B on the screen, and I will read again: "What specific strategies and tools can be implemented to monitor and manage performance and accuracy of a generative AI-enabled device implemented across multiple sites, ensuring consistency and addressing potential regional biases and data variations compared to the device that was authorized?"

Dr. Bhatt: [A panel member sneezes] God bless you. We have several comments that we have made throughout the day with this. So, I'll just tell you where we are already.

The things that have come up as specific strategies to be implemented, especially with regard to regional biases and local variation, are 1) corrective needs post-draft creation when it came to talking about text drafts or interpretations of images; 2) the percentage of misinterpretations at a local institution; 3) the ability to audit using either synthetic data or quality review using locked datasets and varied prompts; 4) or

audits of how many queries or flagged outputs did occur and what areas they were in; 5) the report of an evaluation of the local dataset, and the comparison of what that looks like compared to the training dataset to see whether it's similar; 6) longer-term measurement of high-level shifts that might happen in the answers that we are getting from these generative-AI devices; 7) a measurement-- This was a nice one to have. A measurement of meaningful differences in output rather than non-clinically relevant difference. We don't have a definition for that yet for anyone who wants to give some. 8) The time spent overseeing the output and whether it's increasing or decreasing over time; 9) similar to biosurveillance literature, a look at downstream care-- Again, nice to have. And variability in doctor's judgments and patient outcomes. And then lastly, 10) stochastic sampling, de-identifying images, sending them back to the manufacturer from the local institution for verification and having that run postmarket at a regular interval.

I will go to Dr. Radman now. Oh, sorry. Okay. Do you want to go in between?

Mr. Tazbaz: Yeah, no, that would be-- Thank you. This is Troy Tazbaz with the FDA. And so, the postmarket monitoring, I think there's really two aspects that we have to really consider with generative AI. And often we talk about the more quantitative nature of the monitoring, which is really looking at the performance of an application. And there's the qualitative monitoring, which is what the end user would want to see to ensure that the application actually is performing the way it was marketed to them. And I think that's the part that maybe gets missed a little bit. And I like the providence of understanding the notetaking and whether watermarking it to ensure that it was generative AI-driven generated note. That's a very important one.

So, I guess when we're thinking about the answers to these questions-- And I think there's the technical answer, which is more like a traditional software application monitoring, and then there's the generative AI specific answer, which is completely unique to this problem. And I think-- So, as there's a deliberation happening, I hope that we would probably be able to-- Hopefully we can differentiate those two and highlight the ones that maybe are brand new to all of us.

Dr. Bhatt: Okay. Dr. Radman, take the challenge. Let's see.

Dr. Radman: Okay, generative AI, specific recommendation. Sorry, I'm just going back to a, really quickly, that there should be tiers of postmarket monitoring. And for example, if a device is using a foundation model that they don't control, whether you call it soup or whatever, but that-- A change in that underlying model should trigger a more rigorous review, maybe going all the way back to FDA as opposed to maybe just a local PCCP-type of review where it's like, "Oh yeah, we passed our pre-determined measures and we're still good."

So, that's one idea. But now, going to this idea of regional biases and Question B, the theme of my comment is on user fees, which is going back to-- We talked about how hard it is to change, like, congressional-mandated regulations. And Dr. Dreyer said, "I don't care about that. I don't care if it's hard, we got to do it." User fees. Another thing that is, like, rigorously negotiated between FDA and industry over the course of years, but here we have a paradigm changing course. We're suggesting here where it was such a shift to postmarket, and we've talked about all these new entities. We've already talked about quality-assurance labs. And ArchAI presented. Dr. Dreyer has its AI-health challenge with their abilities to have a conglomerate network validating

devices. And so, all these things cost money. And right now, those people are funded with probably grants or some small-scale funding, but the manufacturers are in this to-- You know, generative AI is a highly lucrative business with probably the most growth of any industry to my understanding. And so, we should consider the user fees for generative-AI products and these new needs that they're bringing up in postmarket monitoring to come with it. You know, additional user fees or however you want to structure that to account for these new entities.

And I'll bring up the example we already have in terms of regional monitoring. So, we already have the ability to conduct decentralized clinical trials, which involves remotely deploying a digital-health device, which may or may not have anything to do with generative AI. And then, to get the data back on the device through a hub-and-spokes type of network that involves a large academic medical center that in turn is able to reach into its community clinics and local hospitals. And they in turn are able to get to the level of community advisory boards and individual community members. And so, "All that costs money" is the point.

And so, I think if we want to be serious about how we're going to be monitoring, and in particular these marginalized communities that don't have grant money coming in and things like that or other sources, we should consider that the user fees for this, the privilege to market your devices, be accounted for those new requirements.

Dr. Kukafka: Rita Kukafka. I have a question that we keep on, I think, going back and forth over the last two days. Is it in the FDA purview to consider outcomes like improved quality of care, decision making, clinical significance? It's a question.

Dr. Fulmer: I would just say that for each device, the intended use matters. Like, the evidence that you have to provide to demonstrate that use is what will determine what kinds of studies we're going to be asking for. And if you're making claims about clinical outcome studies, we would expect that to be demonstrated within the submission as well.

Dr. Bhatt: I think it's an interesting question because-- Well, I'm just going to say it. Because it might be easier to get through by saying "Efficiency" than "Clinical outcomes," and that's not where we want to get. So, I love that you asked that question. I think there's a larger discussion around how we get there, but I love that you asked that question that it's on the record. Dr. Shah.

Dr. Shah: Great. Pratik Shah. I want to spend five seconds defining this thing in machine learning, Tom. And then we'll do the regulatory science framework just to get a sanity check on some of this.

So, when we use the words "Out of distribution" and "Data drift," they mean very different things. An OOD data point is a data point that's very infrequently encountered in the data. So, that could be an underrepresented disease, an underrepresented community member that the data doesn't have. That's technically defined as "Out of distribution data". And then you have a data drift problem, which is very different in machine learning terminology. A data drift problem is that your entire dataset is changing over a period of time. And then if you're looking at GenAI models, then you have the third paradigm that your model is changing. Now, that could be because the data has changed, the model architecture has changed, or you're looking at OOD data that was never there.

So, I just want to define in machine learning terminology what this is. And we can throw any regulatory framework around it. We can ask postmarket surveillance, we can ask all sorts of things. So, standard ways we do it in our research laboratories and even in industry is that we have a reference dataset. And that's a benchmark dataset where in-distribution, out-distribution, OOD, and data drift are defined. And we just run algorithms on that and generate mean and variances. And that's like standard usually. And you have that reference dataset, you have the performance of the model, and then you make any changes that would change. So, now you can put any pragmatic requirements around that reporting. You can say "Don't report it," or whatever the FDA and industry and everybody in ADCOM feels comfortable recommending for that framework. So, that's one piece I wanted to bring up.

The second piece is a very conceptual piece, which is going to the data variations. Historically, if you're building a model on retrospective data, the model was generated using clinical decision-making points that happened in the past. That's what the model has captured. That's why it's called a model. When you use a retrospective-trained model on prospective data, you usually don't have the clinical policy that was used for that model because that's gone, right? So, that's technically-- We define that as "Off-policy machine learning". The on-policy machine learning is you have a policy, you generate a model, and that policy is available in the real world to test the model. That's called "On-policy machine learning". In healthcare, because you cannot use that policy repeatedly, or it's not available, or it's not safe, we call it "Off-policy machine learning".

So, you really are going to look into this retrospective prospective policy, clinical decision-making shift as well as you deploy these GenAI models. So, the data is

changing, OOD is happening, data drift is happening, and you have a conceptual statistical framework where you're doing off-policy machine learning. Essentially, this is off-policy machine learning because it's a model, it's deep learning.

So, I really think that if you are going to give guidance as ADCOM and as FDA, we really need to tell people what we mean by data drift, what we mean by OOD, what you mean by on and off-policy machine learning, and how-- Because the people that are building this model are machine-learning people. How do we address a machine-learning problem with purely regulatory science language? So, we have to create language, literacy, vocabulary, and then taxonomy and nomenclature to meet everybody in the center.

Those are some important points I want to bring up. Thanks.

Dr. Bhatt: Thank you so much. Dr. Soni.

Dr. Soni: So, I'll start by acknowledging that I have to really challenge myself to think very differently about generative AI than predictive AI because naturally, at least for me, it's easy to think about predictive AI, that's helping me with decision support. But I think it's important to also think about when we are looking at this label card or model card, what is it actually telling us? So, often think about whenever we are reading literature on studies, there is Table 1 that describes which are the participants that were included in this study. And similarly, there has been discussion about the need to understand what's the demographics, what's the distribution of the underlying population where the model was trained. But then, if it is being continuously validated, there needs to be comparison between where the model was trained and what's the population where the model is being validated and deployed. And I think that gives us a

population health understanding of the model's performance. But at the point of care, there needs to be a better understanding of how that particular result that's generated from multiple different signals-- How does that map?

And I really appreciate it, Dr. Shah, your explanation about out-of-distribution versus data drift because so much of this by the iterative learning of the model that it's doing is going to have essentially a precision output for that patient in front of the provider or for patients themselves if they're the end user of this. And we need to better understand what's in the model card when it's a personalized result that's being produced for that patient or for that provider.

And I started thinking more about, Dr. Radman, your comment about decentralized studies, that's a lot of things that we have been thinking about, and bringing care to patients' home, and research to patients' home. Well, if there are models that are able to turn a single lead EKG device to now start generating more signal and more information that goes beyond what we see from a single lead EKG output, how are we going to be able to rely on it? And how much of that understanding is needed to have for the safety of what decision is being made? And I do think in that aspect, we also need to not lose track of clinical outcomes. And clinical outcomes need not just be hardcore clinical outcomes. I think there is the triple aim that has expanded to quadruple aim and now quintuple aim of healthcare, of thinking about clinical outcomes, quality of care, patient satisfaction, cost, but then also health equity. And those may be good grounding points for us to evaluate the outcome at a more clinical-significance level. But to get there, that's the most distal. We need to start from comparing Table 1s to each other. Then when we are applying it in clinical practice, what's the significance of the actual value that's generated with the product or with the

output of the generative AI? And how does that inform the clinical decision making? And then think about the clinical outcomes.

Dr. Bhatt: That's great. Thank you. I'm going to take it back to you, Dr. Shah, for one minute.

I really appreciate it when you talk about OOD. And clinically, when I translate that into my language, that's the edge cases that I see, which is not the majority of the diagnosis we make, but those edge cases, which my clinical experience allows me to know.

It makes me start to think about the reports that we're using and we're going to be putting into generative AI, and when that generative AI starts to use generative AI-generated reports. And I was wondering if you could just educate the group a little bit about either early or late model collapse, and whether you think that's relevant in this healthcare arena when, over time, because so many of the reports on whatever chest X-ray you want are going to be more in the normal range than for me, a Congenital Heart Cardiologist, my cases will be less and less available for generative AI to even find.

Maybe, could you teach us a little bit about it? And then, ask us whether there's any role in us addressing it now or if that's just a little farther in the future than where we are.

Dr. Shah: No, I appreciate your question. I think it's a groupthink, that's the way I formulate this. We have to figure this out together.

So, I think you're bringing a very important point. When a model sees an OOD dataset, and if it's a generative model, it could do three things. And I think Dr. Kottler from Stanford, who presented this morning, really gave us a real-world implementation.

What we don't know from her dataset is how many errors that model made, what the data distribution was, training data, if I would say, for that generative model. We don't know the training data, so we don't know how many of those errors they made were OOD errors or data drift errors. So, that's one aspect of that. So, we can figure some of that out.

If a generative model potentially sees an OOD data point, it could do two or three things. It'll create a new data point and a new outcome that wasn't in the training data. It has never seen the left-right, right-left, a man having a woman's diagnosis, gender-based switches. So, that essentially potentially could be happening because it's confabulated or because it's creating synthetic data that doesn't exist, or because it has not seen this example before.

In the healthcare setting, if those kinds of things happen and we don't catch them, we have human clinicians or patients in the loop, and they would catch it and correct it. But if we don't catch it, then it'll compound down into situations where it'll just create more clinical decision making. So, that's more workflow.

About your specific question, about clinicians not having enough expertise on certain cases that they haven't seen in residency and fellowships, which happens all the time-- Dermatology is an example where darker skin tones are rarely encountered by residents and fellows. We wrote a paper on it actually with UCSF. And practicing clinicians can't distinguish dark skin dermatology problems from people who are fair-skinned or have lighter skin tones. If that happens, then the experts in the loop themselves are not able to audit the model.

If you look at-- I'll give you an example of how this works. In elections, if you asked ChatGPT, "Can you recommend what I should do with elections?" It said, "I'm not authorized to give you that answer." I don't know how many of you tried that. If you went to ChatGPT or any LLM and asked a specific election-related question, it would say, "I cannot comment on that."

And I think really giving those model cards or data cards with really good OOD points and asking the model with uncertainty estimations or prompts back saying "I've not seen this before. I don't know what it is," is really how we would fix it on the machine learning side at least. If that helps. Yeah.

Dr. Bhatt: That helps a lot. Thank you. Jagdish, did you have a comment?

Dr. Khubchandani: A quick question maybe for the FDA colleagues here. I'm thinking of the logistics public health aspect planning. What specific strategies do you use to manage the performance?

I have some statements from speakers throughout the two days. "Local problems need local solutions," "The promise and perils of GenAI are context-community dependent in a way," and "There's a need for patient engagement." And finally in the docket there was a comment, "What does FDA see as the role of industry in helping establish local governance and accountability and patient engagement councils, which would maybe help the industry stay more active, engaged and reduce undue burden on FDA?" So, the idea is to create local governance, accountability, patient-engagement mechanisms.

Mr. Tazbaz: So, how do I answer this without sounding too legal?

So, about the localized comment, I think this is kind of going back to Dr. Shah's comment around out-of-distribution. And so, one monitoring capability that I can think of is that when you deploy a model that was trained on a specific dataset, can you ask it specifically around the new dataset that it's being applied to, whether it has the same distribution of population that is? That's a monitoring capability. Now, all of a sudden, you're taking this postmarket monitoring capability, you're applying it to validation of a model in a new environment, which technically is the same thing. And so, I would say that's a capability that we need the industry to think through around its integration of a model into a newer environment. So, now we're talking about multiple site deployments. So, that would be one of them.

Now, about the localized engagement. So, we have specific authorities that we can do around convening. And so, we're always going to be bound by those laws that say what we can do and what we cannot do; however, what we do is we work through collaborative communities and we can maybe develop a framework around what a collaborative community could be but not necessarily execute to that collaborative community. And so, that's one solution that we might be able to address. I don't know, Sonja, if you would like to add anything to that.

Dr. Fulmer: No, I think the collaborative communities is definitely a good topic to raise because I think that's where we see experts come together and come up with some standard approaches and best practices that they could deploy with respect to patient preference, for example. That's something that we do consider when we're making individual device authorization considerations. And if a company, as part of their studies, is able to identify patient-preference information that can help support their marketing authorization, that is something that we're also interested in.

Dr. Radman: I wanted to address this Point B here, strategies and tools, and say that there is a field of research known as "Uncertainty in AI," and I think it's highly relevant to this talk of OOD and add a sample of users or patients. And you can imagine a tool being set up where instead of the model card having maybe course distributions of patients, you could actually have a tool where a patient maybe enters their electronic health record, and then you have this high dimensional description of the patient and you're able to say, "Okay, for this model, your uncertainty is this wide because you're less represented in the training dataset," whereas another patient, "You're more represented, you're more certain device outputs". So, Uncertainty in AI, I think it's just a general field of research worth looking into deeper and perhaps considering developing tools around for these monitoring tools.

If no one else has-- You have your card up. So, I'll stop talking there, and I have some other comments that I could do later.

Dr. Rariy: I actually just had a comment probably similar in moving that thought, so feel free to add.

So, what came to mind was the concept of federated learning. And the way that I've seen it used today is more around training the models. And perhaps there is something that we can take from that when we're looking at monitoring. So, if we're looking at specific tools, the concept of federated learning is an opportunity to take data locally and take model training locally with very specific security and privacy parameters of that local site, and then kind of upload it to the cloud if you will, and allow multiple sites to do the same to one specific cloud. And it's de-identified. And the data is aggregated, if you will, in this de-identified way. And any model updates locally

that occur are also uploaded in this de-identified way. So, it's a process by which you can aggregate de-identified information and continue to learn and continue to train the models where sites may not have the ability to have access to large data.

And perhaps if we apply that same concept but in a different way around monitoring, there is an opportunity for smaller sites, perhaps less sites with less access to resources to still participate in the ongoing monitoring that is required, but provide the possibility that you can have a centralized repository of information that is monitoring on a larger scale and also locally.

I don't know if such a thing exists, but when we're speaking about specific tools and strategies, perhaps it's a concept that we can look to move into the monitoring stage.

Dr. Maddox: That's a great idea. I guess the question-- Maybe Sonja, I'll give it to you, because you answered last time. You said that clinical outcomes are part of the consideration in the premarket stage, right? And is that also true in the postmarketing or where is the purview and scope right now for the FDA in that?

Dr. Fulmer: So, what's key here is that a product is authorized with a specific set of claims. And so, if the claims make clinical outcome claims, then we want to see that data premarket, and we expect it to continue to perform postmarket consistent with its premarket performance.

Dr. Maddox: And you can indicate that you need to demonstrate that empiric data to show continued delivery of those clinical outcomes in the postmarketing period.

Dr. Fulmer: Right. So, there's a few different mechanisms for that postmarket demonstration. For higher risk devices, of course there are more possibilities for

required postmarket approval studies or postmarket surveillance studies that are for specific higher-risk device types. And Jessica Paulsen told us today a little bit about some other mechanisms that we might deploy for low or moderate risk devices including establishing special controls that might say we might need a postmarket performance monitoring plan, which would specify, for example, that we expect to see those claims to be continued to be monitored throughout its performance on the market, or if there's possibility of a predetermined change control plan, which would also likely contain some postmarket monitoring aspect.

Dr. Maddox: Okay, good. So, that's already within your authority?

Dr. Fulmer: It is. Yeah.

Dr. Maddox: Okay, good. And I think that largely just confirms what we're likely all saying about having those be aspects of how we do postmarketing.

Dr. Fulmer: Yeah. And what we're really interested in hearing from you all is particularly when we're looking at one of these plans or looking at a postmarket approval study, what should we be looking for GenAI-based devices to ensure that these in particular will continue to perform as expected?

Dr. Maddox: Right. And so, I think that gets-- And we talked about this a little bit. On clinical outcomes, there will be, I assume there'll be a tug of war between the manufacturer saying, "I want the easiest outcome to measure to get it across the finish line and into the market," and then all of us on the side of care delivery and wanting to make sure that we're ensuring these are as effective and safe-- And to Ami's point, we don't want to just say, "Oh, you're more efficient and that's a sufficient outcome". But I'm also sympathetic to - we can't say, "Oh, do we reduce mortality?" just because the

sort of logic between one decision made with generative AI assistance and everything that will go into somebody living or dying is actually quite far and a little unrealistic.

So, it does make me wonder. And we were talking about this a little bit with the radiology presentation, perhaps the governing concept on an appropriate outcome that may not be too far reaching is the accuracy of the decision being made with that generative AI input. And it gets back to that radiology example. I'm sorry, that has clinical import. And so, it gets back to that radiology example of "Can we rule out a pulmonary embolism in this patient?" because that kind of outcome is going to be from a time point of view right then, but has extreme clinical consequences. Somebody gets home from the ED and is fine, versus you missed it and now they're going to go home and have real health problems.

So, we probably have to play around and make sure that concept is applicable to the broad number of use cases that eventually will come under consideration, but I wonder if that might be a balance point between the clinical outcomes, we really want to protect without making it so unrealistic that it just sorts of mucks up the process.

Mr. Posnack: Steve Posnack. Oh, got a little loud there.

So, I'm an optimistic heart and I have faith that this will get better over time, which I think from just regulatory experience that we do see that what seems valuable would be to discuss how to approach this over time in terms of what can be incrementally done. And so, generative AI is knocking on the door, and I don't think we want to be talking about how to do this in 2030. So, we have a little bit of a challenge of what's going to be most effective now versus later over time, which I'm sure will get better.

And so, "What can be used now?" I think is an important question for us to ask ourselves and help weigh in for our FDA colleagues about how to prioritize, what can be used now, what's the best for 2025 and 2026. Because we know it's going to get better, we'll get more insights. I think we saw from some of the earlier presentations today that the industry will learn as well other metrics that will be more impactful. And then, there are other systemic changes like user fees and other types of negotiations that will just take a lot more time. And so those aren't accessible in the near term to guide this work and to help start the process of getting familiarity in industry.

So, I think those are things that we're going to have to figure out how to balance best.

Dr. Kukafka: Just two quick comments. One is I don't think we can with certainty assume it's going to be around for a long time because, I mean, I've seen AI come and go over the last 20 years. So, the fact that we have it, if we don't do it right and we don't-- That's why I keep on thinking about-- And you did talk about outcomes. I mean, mortality I agree is not realistic, it's not realistic in non-AI, but other types of more immediate endpoints might help us ensure that it's going to be around for a long time.

So, I just to make– I just want to-- AI is not the first time it has come and gone. So, I just think we have to be careful. And I think we have to be careful with respect to showing-- And you brought up the quadruple aims. If we could say that it's more than useful, but it impacts health, safety, quality, short-term outcomes-- I just think that if it is in the purview of FDA, I don't know, but I think it will help ensure that it's not going to come and go. Just a comment.

Dr. Elkin: So, I think for the short-term and a long-term view of this, that we obviously want to set up registries where people have mandatory reporting of, as we said before, both random sampling from the population that they're using the AI on, and also allow for reporting of adverse events that we keep track of. But I think it would also be appropriate of a Committee like this, even though this isn't simply a recommendation to the FDA, it's a recommendation to the government that we do set up a medium, where we're tracking longitudinal data from everybody all the time. You can imagine in the future if we were tracking such longitudinal data that one could run studies in the cloud that were both national, based on demographics, and also regional. And you could do this over and over again at periodic times.

Now, the idea of the cost of this policy and regulation is something that comes immediately to mind. So, if you are making every local place do this on their own, it's going to be very high cost.

And also, you have to think about a little bit what is the cost of running these models for the additional workload that it would have to do the monitoring that we're talking about here too.

So, I think it is reasonable for us to consider in our advice to the FDA, what is really-- One is what is really essential, and two is what is an acceptable burden given the cost structure in the United States. And this part has to do with who's going to be paying for it. But assuming it's device companies now and looking at the market that exists today and the potential market that exists tomorrow, how can we plan for a schema that is going to support the industry and also support our patients and their families equally?

So, I think there needs to be some more work on this. I don't think I have all the answers, but I'm absolutely certain that if we had an infrastructure in the country that had all the data in one place that we could use for testing-- So, if you put a new change to the model, you could go back and use it against retrospective data where you already have the outcomes and then you get validation and it should be very fast to run, easy to do and relatively low cost. If you have to build this for every single study that you're going to do, it's clearly prohibitive.

Ms. Miller: Thank you. I'm going to be a little shorter this time. I just want to follow up Dr. Paulsen about what we can do today and bound us a little bit. And I'm going to play words here. We need to have an adaptive framework for these adaptive algorithms. So, we're probably going to need to iterate. I know that's not easy, but we're not going to know it all. So, we have to start somewhere and then iterate over it. And I think we've been doing something like that recently, but I think we have to start and keep building it when we learn.

And going back to the cost and the burden, I'm going to say something maybe a little bit controversial, but who gets the burden? The burden is not going to go away. There's a burden on the user of the device to validate the output if we want to go that way. There's a burden on manufacturing to put extra controls there. There's a burden on FDA and the payers. So, I think we have to spread that, and it is going to happen if we want to add the extra controls for these devices.

And what made me think earlier about that example from radiology, how burdensome it is for some of them to go over these cases and see if it's true or not true.

And I'm not sure if we want to have humans validate this forever. So, something to think about.

Another point about the cost, the adoption of this technology. And we talked about the codes earlier and the payment and the reimbursement and all of that. So, the adoption is not going to jump up if we don't solve the coding and the payment and reimbursement and all of that. So, yeah.

Dr. Bhatt: Troy, over to you.

Mr. Tazbaz: Thank you. Troy Tazbaz with FDA. One, I think that the payment conversation is probably out of scope of this dialogue.

But a couple of things that I want to talk about are statutory authorities. So, although we do have some elements of postmarket monitoring, authorities, generally speaking, they're driven by adverse events. And so, I think when we're thinking about additional authorities in order to truly do this the right way, then you have to have more proactive authorities to be able to monitor performance before these models start failing, not after they failed because the adverse event at that point is much greater in many ways if you deploy this at scale. So, I think that's one issue we have to tackle.

And from a capability perspective, and I think Diana said this very well, PCCPs are about a roadmap. It's about documenting the roadmap of the product that you're going to be deploying and what additional features you may be adding to it without changing the intended use or any optimization that you might be able to drive through new information that is coming in that allows you to retrain the model itself for just maybe better performance. Well, this is kind of analogous to that where what we're asking today is we're not going to be able to solve all of the monitoring capabilities that

are required probably by 2030 when maybe there's a higher adoption and usage of these things. But in a sense that what we're really trying to get out of this particular conversation is almost like there's an FDA element of what we should be looking at from a capability perspective. But it's almost like a product requirements signal that we're sending out to the industry around. Where do you start? What are the critical capabilities that you have to address in the product monitoring capabilities that we will get to add?

And so, I've been in the kind of technology industry since the mid-90s and I remember building my own monitoring tools using Perl, and Shell, and scripts back in the 90s and then all of a sudden, the industry took over because it became critical. And then, once Cloud really came in, particularly with software as a service, they started building all that capability directly into the product itself where I could see this also going too. But in the end, we're still back in the early days of the 90s when it comes to this particular capability.

So, what I would really like to get out of this conversation is "What are the absolutely most important critical things that we have to do today in order to start perhaps integrating these things into a critical care practice today?"

And then, I think that as we learn more and more about these capabilities and of course the application expands or extends, then additional capabilities will have to also be implemented when it comes to monitoring.

Dr. Radman: Thank you. I just wanted to just state that guardrails are a topic that I feel like has been a little bit neglected in this meeting. We really haven't had an expert

speaker talking about the use of guardrails to continue to ensure the proper usage of a GenAI model in the field.

So, Dr. Shah gave the example of the election prompt, and that's a guardrail that was put up after biases were shown in that query, right? So, there's this-- I think it comes down to the PCCP performance monitoring plan idea where what the guardrails that have in place are-- Or not even what the guardrails are because you don't know the guardrails-- They didn't know they were going to need an election guardrail around election time. But what is your process of putting up guardrails, or detecting the need for guardrails, putting them up, validating their effectiveness?

And yeah, I think it's a really broad and important topic. I'm not an expert in the topic, I'm sorry, but that would've been-- I think if anyone here has more information on that, I'd appreciate it at least.

Dr. Rariy: Dr. Radman, I feel like I continue to expand on the thoughts that you've laid down. So, thank you.

One of the things I wanted to specify is the concept around protocols for adverse events or protocols of error reporting and creating that in a mitigation strategy. So, as an example, Troy, you spoke about the-- What I took away was the definition perhaps of error needs to be defined in generative AI as it's different from what we would consider other areas. The definition of adverse event needs to be defined specifically. And perhaps that's a role that the FDA can play or at least keep it specific but broad in that the definition is left up to interpretation to a certain extent because we don't really know what the error definition would be 10 years from now. But I do think we need to create some parameters or some guardrails around what that is.

As an example, in the example that you gave, Dr. Radman, around the election, whatever was the prompt for ChatGPT to create those guardrails around that election, that would be how we classify, as an example, an adverse event or an error. So, there's some trigger by definition that would warrant the device company to then implement some sort of strategy. And perhaps, as the FDA, we could articulate a need to have a process in place. What is that device manufacturer's process that they're going to monitor and adhere to for various errors that come up? And part of that could include or should include a very specific mechanism by which reporting can occur.

So, we heard as an example from the patient advocate that who to call and what to do when an error is reported is not clearly understood. And so, that's part of the necessity of that device company to make it very clear to the end user how to report back specific errors that do occur by this particular definition.

The other is just to create, again, a framework around the error and adverse event. I think they can be two separate things that need to be clearly defined. The type of error, the type of severity would need to be included.

Dr. Radman: I was told I was allowed to riff off of Dr. Rariy's comments, and just to build on it.

You could also imagine like we were talking about feedback loops of patients or other mechanisms of reporting errors. And if you did have a crowdsourced system like that, you'd get some that are true errors and some that don't rise to the level of an error to act on as per the definition of adverse event here that Troy mentioned. But you could also envision having some adjudication board, like, we have a data safety monitoring board, some kind of error adjudication board that does review those or the ones that are

most frequently reported, for example. And "Okay, this is a ground truth gold standard error," and then you could train machine learning algorithms to crunch through those databases and start raising things up faster and more frequently.

So, that's a really kind of synergistic loop there to have also the data being part of the process here.

Dr. Bhatt: Fantastic. I'm going to take a moment to tell us how far we've come just so that we can then continue talking.

For postmarket performance monitoring, we said "What specific monitoring capabilities should be considered?" We have three groups that we have talked about. The first is Human and AI Interaction. 1) There should be a statement in this postmarketing period where there must be significant human oversight or the device doesn't work. 2) There should be watermarking across everyone. So, we know that the use of generative AI was included and it's transparent. 3) And we need to use existing data standards and add to those standards to work with CTA and others what GenAI-enabled devices might need us to add for those standards. 4) People are asking for a low-cost monitoring mechanism at a larger level, whether we call that registry or reporting. 5) We'd like people to explain which benchmark datasets were used because then in the long-term postevaluation, they can be used against those same datasets. 6) And then we need facilitation of error reporting to the manufacturer and the end users, both clinicians and patients. Again, going back to perhaps a low-cost, low-effort central database. So, that was Human and AI Interaction.

The second area that came up is Usability and Risk Mitigation. So, starting with adverse event reporting, perhaps 1) Accuracy: Did it tell you what it was supposed to?

2) Safety: Did someone get hurt? And 3) Bounded use: Did it give you an answer that is so far off from what the intention was? As a starting point for adverse event reporting. 4) The likelihood of the user to find the error reporting goes to usability and safety. 5) And then, internal periodic review of whatever it might be, chat transcripts as the example, documenting whether there was product failure or implementation failure. And then defining for the groups what OOD is, what data drift is, when changes in the model occur. So, education about nomenclature being necessary for us to actually be able to report risk, otherwise we can't report risk. And then resending to the FDA if the foundation model massively changes that you're basing it on.

And the third group is Thinking About Using FDA Models for Baseline Infrastructure. We separate Product PCCP from Use PCCP. We're going to focus on Product because that's what we've done today. Thinking about description of modifications that are coming, globally or locally implemented, making sure the device is maintained with an intended use for now, according to what we've heard today, having a modification protocol, and ideally using the PCCP Impact Assessment as perhaps an interval necessary protocol postdeployment for a period of time because we already have a way, as our industry colleagues mentioned, to measure people in the postmarketing period. Perhaps we can use that. We'll add to that, conversation about the TAP program from yesterday, and perhaps adding clinical informaticists and clinical implementation advice both in the beginning premarket, but could that same TAP program be almost like the DSMB who meets in the postmarket setting and says, "Did the things we thought were going to happen happened? What was different?" such that we don't have to stand up an entirely new organization just yet?

For B, we have shorter answers. Specific strategies and tools for regional biases. 1) One is data: report evaluation of local dataset versus training dataset. You're using something that was trained on this, does your data look like it too? 2) Ask for ensembling, which we learned about earlier. Embedded quality insurance checks that we ask the device manufacturers to put in, and that can then be automatically reported. Error assessment, corrective needs percentage and types of misinterpretations, OOD, data drift measurement. 3) Auditing: whether it's synthetic data, quality review, queries of flagged outputs, or stochastic sampling. And that can all run automatically. And then, 4) long-term measurement, high-level shifts, downstream care changes, patient outcomes, time spent using these devices, and then practice changes over time.

That's where we've come so far between A and B. Am I missing anything significant thus far?

Let us show C. And then we have Dr. Soni, Jagdish, Peter, and Dr. Shah. And I apologize for going between first name and last name. I'm just a little tired.

Mr. Tazbaz: Yeah, and I think you covered a lot of the points also for C as well around the multi-layer application design. But this is a very interesting area that I would love to hear from the Committee.

Dr. Soni: So, I'll make a quick comment regarding outcomes and maybe in defense of efficiency. And then, I would ask a question from the FDA about methodology that relates to C and potentially some of the premarket process.

I do think it's important to delineate between efficiency and it's not necessarily volume, but it's also what it means. I think it's a good thing if efficiency can be defined as more time between patient and provider to talk to each other. I also think it's a good

thing if it takes less than 15 to 20 clicks, which is one of the recent reports showed to prescribe ibuprofen when a patient is in emergency room. If there is a generated automated order based on all of the data that has come in that makes it a single click, it's a good efficiency example.

So, I just want to make sure that when we are looking at outcomes, we're looking at outcomes more holistically and thinking about the art of medicine and not just the hard outcomes of mortality endpoints.

The question I have for FDA is: What appetite is there for-- We kind of brought up the idea of comparative effectiveness, pragmatic studies, which I think there is a robust literature surrounding ways that it is similar to some of the more rigorous randomized control trials, and recognizing that all models are wrong, some can be useful, how can we leverage methods that in rare disease approval have been used? Synthetic control trials have led to approval for therapeutics in that space. If we are able to demonstrate that it's just very hard, both from a timing and pragmatics perspective, to do a full blinded randomized control trial, is there appetite for synthetic control trials or target emulation studies to be able to generate the effectiveness data? Looking at different outcomes in the FDA to consider that.

Mr. Tazbaz: …To help address this question.

Dr. Soni: And perhaps, Dr. Paulsen, both synthetic data but then also synthetic controls.

Ms. Paulsen: Yes, I don't know how deep we'll be able to get into it, but I do think the use of synthetic data is something we will consider. Obviously, the application matters, so the specifics matter. So, for anyone thinking about doing it, know that we're open-

minded, come and talk to us about a proposal and we'll make sure that the right study gets designed to give us the evidence we need to be able to, at the end of the day, determine that the device is going to be safe and effective. So, thank you for your question.

Dr. Soni: Sorry if I-- Just a quick follow up. Are there bounds to what outcomes can be considered with those designs or other outcomes, more patient-facing outcomes or some of the process outcomes? Are those also considered for that framework?

Ms. Paulsen: Yeah, I mean, I think it's difficult to say broadly. I think it's going to depend again on the device, its intended use, and what the sponsor intends to claim that that product can do for patients or providers.

Dr. Bhatt: Dr. Khubchandani and Dr. Elkin.

Dr. Khubchandani: Yeah, I just have a quick-- Jagdish Khubchandani. We talk a lot about data that will be collected, needed volume, velocity, variety and all-- Like Dr. Elkin mentioned, registries. I would like to see some more discussion on the Data Sharing Plan. I hope the community researchers don't have to pay thousands of dollars to buy the data, or the patient has access to aggregate data. Clinicians have access to what's happening in this postmarket surveillance data. So, I would want to see more data sharing transparency built upon that. Thank you.

Dr. Elkin: I actually wanted to talk about that. So, in following up the whole last discussion, we also have to consider real-world evidence trials that have already been identified as a goal of the FDA. So, this is in view trying not to reinvent the wheel. I want to make the panel aware I serve on the NQF Expert Panel for the common formats, which creates the formats that are reported to AHRQ for patient safety to the patient

safety organizations. And it's all around the Patient Safety Act of 2005. And I published a great paper on it if anybody wants to read it. But the idea is that we have a device format already that's been in wide use, and is out there. And it already has elements of AI and GenAI in it. Is it good enough for this panel? Maybe not. We'll have to look at it, but we shouldn't start from scratch. We should use what's out there. And there's also patient safety organizations which have mechanisms for gathering that data already. And there's a data warehouse at AHRQ that was generated by this that is accessible by both academics and patients and community groups and so forth. So, at least we should be aware of that as we move forward.

Dr. Shah: Pratik Shah. I'm going to answer Sonja. And Troy, I'll try to answer your burning questions the best I can. And pardon me if I'm not speaking regulatory legally. I'll try. You help me out here.

So, usually when you're going to ask methods and metrics-- Like, uncertainty estimations, you guys are already aware of and I'm really glad you're thinking about it. It's so forward thinking of you to think about that. It's an emerging field. So, standard uncertainty estimations are done the following way: You do bootstrapping, which is you take the entire data and you export all segments of the data and generate error bounds. And you can do it by important sampling, other methods-- I can talk technical with you later. You can also do model-based uncertainty where essentially you can drop the weights of the model as it's training so it doesn't see sections of your data, and develops uncertainties on those datasets. So, that gives you that if certain populations dropped out, what would happen? What would happen if certain populations just dropped out of the model performance metrics? What would happen? So, that takes care of that.

And the logical way that we do uncertainty estimations and people in randomized clinical trials have been doing it, and we do it in machine learning too, is inverse propensity treatment. Wait, I'll tell you. Essentially you don't treat a model that can treat the entire world like the way it has seen it. You essentially take cohorts of patients and divide them into subpopulations, match them by age, gender, comorbidity, ethnicity, and then evaluate model performance on those specific subpopulations. And in the data output you report, how does this model perform for the indications that were indicated on these specific subpopulations? And you can do that with retrospective data by IPTW and other methods. You don't have to go and recruit patients and do a whole trial if you have retrospective data, but we cannot take technology to people. We have to look at people and then design the technology. That's how we do our research as well. So, we really have to look at the population indications on which these models will be tested, and then develop all these uncertainty estimations, and that-- At least if you decide to put that in your guidance documents-- I'm not saying you should, but if you decide to do it, you will really have a good scaffolding of what the performance metrics on what subpopulation, what uncertainty estimations, what upper and lower bounds in the premarket authorization. And then, as Sonja said, you can then make people accountable that-- "You said that. So, now give us the postmarket analysis of that."

Dr. Bhatt: Dr. Jackson and then Mr. Tazbaz.

Dr. Jackson: Thank you. I'm going to ask you all to bear with me because I've been forming this thought as I've been listening to the conversations based on some of what has been shared. I'm not sure if this goes under A, B or C, but I think it speaks to your question about what needs to be priority, what we need to do now. And I think one of the things that comes up for me that we haven't necessarily deep-dived on is thinking

about the impact for GenAI on health equity compared to software as regular medical devices because GenAI will be creating new data. And I know that we've talked about clinical trials and different ways to capture this, and I hate to say it, but historically we have not been equitable including data from marginalized communities and clinical trials. And we have felt that that was still good enough. That needs to change for GenAI because we will be training future models on the data that is coming from these, and the risk of harm from that for misinterpretation of symptoms thinking about if a model gets trained on some sort of skewed data because of text input, for example. And then, we get to a point where we're saying, "Okay, we don't see this in this specific patient," a black male that's 55, and then later we miss something. How many more people will be impacted by that?

And so, I think even when President Biden had his executive order, that was one of the things he also highlighted. It was health inequities in healthcare applications and how that needs to be scrutinized. And I actually think that that is an absolute need-to-have. And I think that companies often when we're building things, we almost think of it as a nice-to-have and adding it later when we're talking about iterations, I think that actually needs to be a metric currently making sure that the device in its intended use performs postmarket equitably across different communities, that we don't just say it performs at 80%, but it's actually 20% for some communities and a hundred for others, and we get to 80%.

So, I think that is not a fully-formed thought, but something I do want to make sure that we're really thinking about, in the importance of this more important than we've been thinking about it for other medical devices.

Dr. Bhatt: So, maybe a quick question in this direction then to Dr. Shah after Mr. Tazbaz: "In academia for journal publications for example, many journals are now saying 'If you are publishing on AI or algorithms and you do not have a diverse training set, we are not accepting your research letter.' Is that something that the FDA has done, can do, has thought about? Is that within purview?" I'm just curious as to how close to that we can come at the government level, not the academic level.

Mr. Tazbaz: So, I'm going to have to answer that in maybe two parts. And so, we always go back to the intended use, which is: What is the intended use and who are you using the product on? And I think that drives a lot of product-development decisions and of course regulatory decisions. And I think-- I can't remember, someone said yesterday. If I'm going to develop something for, let's say, maternity, well, I'm going to sexually remove any data that is not relevant to that. So, there's always the intended use that has to come into play.

Now, the second part of that question is-- To Dr. Jackson's point around, it's not a nice-to-have, it's inaccessibility to data is really the bigger problem. It's not their desire not to want to have data. It's the inability to actually access the data because of the fragmented nature of healthcare data is a bigger problem in the United States.

And so, this is actually maybe an answer to Dr. Jackson, Dr. Shah, Dr. Elkin, and I would say everyone on this side of the table who talked about this concept of data. So, a bit about my background. I used to design data centers for a living. So, I used to build storage and compute. And I can tell you that healthcare data is about 30% of the global's data creation. And there's not a single entity that is going to want to take on the capital liability of building something that's at that scale from just aggregation. And so,

this concept of federation comes in very important on this one, which is being able to access data at its origin.

And so, I really liked what Dr. Longhurst said from UCSD around this AI implementation science center, which is what we always kind of imagined, which is an access to data in its original location. We tried to do that within a partnership with VA. We've done a couple of projects, and now we've expanded that. Initially it was to test-- Well, to have people develop products and test it on the VA datasets. That was a very successful collaboration. And now, we've actually expanded that to how we actually ensure the access to that data from the outside world while having these specific controls, like intellectual property controls, data privacy controls, and all of that. And that's the idea that we want to bring to the industry itself. But that means that the industry has to be okay with developing the technical design of access to that data while also not necessarily thinking about the concept of monetization of accessing that data. And I think that's a very important part of the equation here.

And so, everything that you all are saying is absolutely necessary for this thing to work from a health-equity perspective, from a performance perspective, from a quality-management perspective, not just in the premarket, but also in the postmarket element of that. But that means that there has to be a fundamental shift in this concept of enabling access to data.

Now, where it gets very challenging, Dr. Jackson, is that you have very valuable data in maybe community centers that have no access to it. So, that means that we have to collectively think about where we aggregate that data, how we aggregate that data to

ensure that there's a representation of that data when AI models are being trained on and qualified against.

I'm not sure we're going to solve that problem here because that's actually not a technical problem. I can tell you that. It's a fundamental problem the way that we're thinking about these problems. And so, unless we can correct that, the technical solution is actually-- If you can get around the financial issues, the technical solution is not that hard.

Dr. Radman: Is it like "Data is a utility," some people say? Or should we--

Mr. Tazbaz: You could use that analogy, that data-- Well, in this case it is a utility, right?

Dr. Radman: Right. But ideally it would be a utility. Like, we all have access to water, we should all have access to community level.

Mr. Tazbaz: Right, right.

Dr. Bhatt: Dr. Radman's comment, just for the transcriptionist, was on whether that's calling data a utility and enabling us to be able to get that information.

Dr. Jackson, Dr. Shah, Dr. Soni, and Miller. And then we're getting close to closing. So, if anybody has final thoughts after that, get ready.

Dr. Jackson: I just want to quickly respond to what you were saying, Troy. And I want to push back a little bit. I don't know if we'll solve it today. I do think it does need to be a part of the conversations even now when we talk about iterations for the postmarket performance because I think it is solvable, the inaccessibility to data. I mean, look at us now. We are literally today talking about people who have figured out how to use

algorithms to create GenAI to create new data. And we're still saying that we cannot figure out how to get access to the data in the communities. And I think that is one of the things that keeps us stuck. People are saying "This is too big of a problem," and we keep kicking the can down the road. And then, what happens when we have this data that is in creating new data because we didn't spend the time solving it? What does that look like for health inequity and the devices that we're approving 10 and 20 years from now?

So, I'm not saying it's necessarily the FDA to solve it, but I do think it needs to be a part of the considerations that companies do need to be held accountable for thinking about it when they're looking at the postmarket because we have the capability to solve it. I mean, the things that we're building now, we can figure out inaccessibility. And I think that is concerning to me because it also goes to the point you said about monetization. That is something we haven't talked a lot about. But that is concerning to me too because everything we've talked about today is "Data is king", and at some point, people might build devices or try to get them authorized to have data to build something down the road that's not helpful for clinical outcomes, and communities can be harmed by that. And I think forcing-- "Forcing" is probably not the best word. Holding accountable companies to have that health equity piece I think will help to decrease people who are just collecting data to have data that might monetize it and further harm these communities.

Mr. Tazbaz: I do need to address this. So, I didn't say that it was a challenge that we shouldn't be addressing. I said it's not a technical challenge. That is what I said. It's technically actually a quite easy solution that we can do, but there're fundamental issues

around the confidence that people have around how data is going to be used. So, there's a privacy element to it too.

Now, "Data is king" comment. The World Economic Forum actually published this study-- And I love the study because it actually helps with my narrative. And those are the best types of studies, right? Right, exactly. That 97% of the data that is generated in healthcare settings never gets used, which means that we say "Data is king," or I could say it's a cost center if we're not leveraging that data. That was according to a study. And I think that that's where we have to challenge the conventional wisdom around what we are going to do with this data, how we can actually leverage that data beyond a point in time that data was generated and used for a specific care decision.

Dr. Bhatt: I would perhaps suggest, based on Dr. Jackson's comments, that rather than a need-to-have, is there a question in the premarket and the postmarket use of generative AI that says, "What did you do and where did you fall down in trying to get a more diverse dataset? And then what is your plan in postmarket surveillance to ensure that this gets to the people who need it? Or do you know how to do that?" And perhaps something like that at least starts to bring the conversation to every company that is coming through even though we're not mandating just yet.

Dr. Shah, you had a comment, but I know Dr. Radman wanted to comment on this topic. I don't know if yours was also on this or--

Dr. Shah: It's kind of related.

Dr. Bhatt: Okay, so maybe Shah, and then Radman.

Dr. Shah: Yeah, I really appreciate what Dr. Jackson said. And me, as a Machine Learning Model Builder myself, and we do that, we always have this existential crisis.

How diverse is the model? And then, the question is: How diverse is the data? And then, the question is: Where do we get the data? And how long can we wait to develop this model? So, I think we have to be pragmatic and industry-friendly with industry already-- I'm not saying that we should just let people do whatever they want, but we also need to create a trusting framework between regulators and industry so that industry feels comfortable disclosing to the FDA that what populations, what they were able to recruit or they already had that prospective data on and where this model has been tested on. And if you can just get that honest assessment that these are the ethnicities, race, demographics, zip codes-- Don't forget geography. People have a lot of diversity in where they live. And we have access to that data and we have tested it, that's like half of the battle one. And if it's a small device manufacturer who doesn't have the capability to launch a trial to recruit people, then we need to help them connect them to data lakes that are federally available so they can at least augment the model performance and give you some synthetic outputs from them.

I think we should really take the role of enabling the small device makers industry to get access to data and help them figure this out. And then, if it's a really large organization which has the capabilities, they can actually do a trial. They can recruit people because that's what the drug manufacturers do. We asked people to do RCTs before we approve a drug. So, why make those concessions because it's a model, right? Because it's still impacting people. So, that was my comment. Thanks.

Dr. Bhatt: Great comment. Going back to the TAP-type program, and then to Dr. Radman.

Not just clinical informaticists, but maybe there needs to be a health equity lens and availability of somebody to talk about how you get a diverse dataset and offer that to the companies who come through our programs.

Dr. Radman and then Dr. Rariy, I think you have a comment on this as well. Then we'll go to Ms. Miller.

Dr. Fulmer: If I could interrupt just really fast, sorry, just to reiterate a couple of the points that have been made so far. I do want to emphasize that a diverse dataset is an important part of our review today for devices in general, for AI-enabled devices and in the future for GenAI-enabled devices. We have denied products in the past based on inadequate performance or at least representation in performance data for different subgroups. We do ask for subgroup analysis when we're doing our reviews. It is something that we do take seriously today as well.

Dr. Radman: Thank you. I'm going to add to Dr. Shah's list "Social Determinants of Health" as an important subgroup. But yeah, I think this discussion, to bring it back to the regulatory framework-- And I think you know this because I agree, I know you do or you're starting more and more to look at performance in subgroups as per we're discussing. So, I do appreciate that, but I think it comes down to, as pretty much everything we're talking about here, to the claim. So, if they're claiming broad use, then there is that regulatory, then it becomes the regulatory purview for them to demonstrate in subgroups.

Now, in bringing it back to this Point C that's on the board here about multilayer application, really talking about foundation models in a way-- We already saw, I think in the talk yesterday, about these biases that we already know to be in these models.

And I'm sure we've all heard about them. "Show me an image of–" What was it, a doctor or whatever caring for white people? Was that what it was? Or something like that, right? And it had it backwards. And there's other examples like that. "Show me a picture of Americans" is one I've seen where there's a bias in the outputs. And maybe those are gone already because of guardrails, right? But then, now we're in a situation of the example of the election response where it's like a blank spot or a blind spot, and that's not good enough to just throw up a guardrail. "Oh, it has a bias."

So, I think if the FDA's purview is claims, then I think you should do everything under your regulatory power to hold manufacturers to performance that fully embraces the scope of their claims. And in particular, with the foundation models that are just broadly creating these devices and they're scraping an unknowable amount and sources of data in a lot of ways, they're harder to regulate. I think that's the scarier part than a smaller manufacturer who's maybe fine tuning. We can know what their fine-tune dataset is if they're building on a foundation model or building something from scratch that's in a focused demographic.

So, it's a major challenge how to handle foundation models and in particular their bias. But I guess I propose some kind of certification program. Like, we've talked about the manufacturer certification program that may or may not be under your purview. But there's also a biomarker certification program. So, I think we could learn from that as well where there's an advantage to be prevalidated in some way and perhaps foundation models may find a place in those regulatory tools, and that can address some of these questions on bias that we're discussing now. Thank you.

Dr. Bhatt: Great, Dr. Rariy, Ms. Miller, Dr. Soni, and then we'll be up against the four o'clock time.

Dr. Rariy: Thanks so much. I think-- Dr. Jackson, thanks for bringing this-- Highlighting this more, I would say.

I think one of the things I wanted to mention and perhaps explain a little bit more is the concept of postmarketing metrics, which would include, and that would be my recommendation, which would include fairness metrics and bias metrics, or equity metrics as an example.

So, to answer your question, Dr. Jackson, clearly articulating that this needs to be considered in the postmarketing and premarketing stage I think is important. And by that, by fairness metrics, it's an opportunity to measure and report fairness metrics based on your specific use case, what that would look like, and to allow the device manufacturer to think about what specific outputs would be classified as fair in various populations or what have you-- Depending on the specific use case for the device.

Similarly, when we're looking at equity monitoring, both on the local level and on the population-based level, I think it would be appropriate to look at what specific outputs are generated across what specific demographic group. And again, based on the use case that might change or be influenced.

Similarly, the third metric would be any identification of known biases of the model, based on population-based specifics, race, ethnicity, rurality or others. But just to clearly articulate that upfront and perhaps the percentage, we can have some sort of ability to quantify what that looks like and any mitigation strategies that would need to be put into place upfront, or ongoing model reorganization that would need to happen.

I think the exciting concept around generative AI, in addition to absolutely the lack of the data from diverse populations that the generative AI is being trained on, absolutely carries the risk of exacerbating worsening disparities.

On the other hand, I think we need to facilitate an opportunity to, one, recognize that the underlying data that it might be trained on is biased by that clear articulation, and if there's not an appropriate strategy to mitigate that upfront and obviously not cause significant harm as well. So, we look at the risk benefit ratio. But then, at the postmonitoring stage, that is one of the opportunities with generative AI is that the amount of input is infinite. And so, we can then take in data from diverse populations just as long as there is an expectation that the model, the algorithm, etcetera, will be changed and updated and evaluated for these changes or drifts specific to this diverse data.

So, I think there's a way that we could incentivize an opportunity to collect real-world data from real diverse patients, again, depending on the use case, and ensure that the postmarketing or the postmonitoring is sufficient enough to take that into consideration and improve upon the initial model.

Ms. Miller: Thank you. Diana Miller here. I just want to bring up a few points about this data access and usage. And I think Mr.. Tazbaz is probably familiar with it. There are a lot of privacy laws, and access laws, and regulatory laws, and state by state they are different. And pediatric laws are very strict in getting data for pediatrics and stuff. So, this is a bigger problem I think for this Committee. And we have to take that into account because we can't have access to data for specific groups.

Dr. Shah: Don't forget HIPAA.

Ms. Miller: Yeah, HIPAA. HIPAA is a big one. And we talk about some HIPAA opt-in mechanisms and all of that. So, it's not that people don't want to, it's a matter of legality.

Dr. Soni: I wanted to comment on the use of federated datasets and maybe go back to the first question I asked yesterday, and see if I can land the seed that was planted, which is surrounding the certified AI assurance labs or AI assurance labs that may be able to leverage certain levels of federated datasets that potentially we can rate them to quantify and make sure that they meet minimum standards of representativeness. And if there are certain, not-for-profit, but certain certified AI assurance labs, much like how food products go through their nutrition labeling and that kind of evaluation where there would be certain standards applied to them as well as certain level of accountability, that perhaps can help with the independent verification and validation within the total product life cycle.

The one part I would want to emphasize in that process is with federated datasets, there tends to be a tendency of data going out, but not necessarily insights coming back in. So, we need to be also thoughtful about, beyond model development, if model validation is performed on federated datasets, what are mechanisms by which some of those outputs and results of those performances can go back to the community hospital, to the academic medical centers? Because that doesn't necessarily happen.

There are existing datasets, like PCORNET and N3C and many others, that could be leveraged for this, but there needs to be a better mechanism to bring some of that data and insights back to the hospitals that are contributing to that.

And I think if we apply this kind of framework, details need to be worked out, we can mimic what Dr. Elkin mentioned, which is large quantities of data, many simulations, many attempts at trying to break the models and see where the model stops performing fairly.

Dr. Bhatt: Great. I will turn this over to Mr.. Tazbaz, and then I'll say some concluding remarks as well.

I'm so happy though that you brought that up again because I think what we're getting at is that there are likely in-device mechanisms to be able to embed the ability to do quality checks. There are system-level or clinic-level mechanisms that we can have and that they're really-- And I'm glad you said the word "not-for-profit". I think the idea of registries or large places where we can put data and bring it back, but not for the sake of profiting off of it, but for the sake of our patients, I think that is the emphasis that we're all getting to here in this postmarket surveillance era, which I love.

Maybe Troy over to you for any final questions that your team might have and I can do concluding remarks.

Mr. Tazbaz: So, we do not have any final questions. This was a wonderful discussion. I think that it was very clear in kind of where we are and what we have to go accomplish. So, that's why we didn't really have any clarifying questions on that.

I just want to maybe finalize by just saying thank you to all. And I'll start with the FDA team behind me and many people who are not here, but they're the ones who envisioned this and planned this, and then of course had to execute it in order for us to have this. That includes the selection to make sure that we're having this type of discussion. We can't also forget the AV team that ran this thing as smoothly as we did.

And of course, the presenters that we had, I think they really did set the stage and a lot of the dialogue that came out of it. And I particularly really appreciated the both questions and answers that were going back and forth.

And of course, lastly, I mean, you all. I mean, this couldn't have probably happened at all if you were not so engaged, and not just interested, but also, you're depending on what comes out of this as clinicians, as people who are in the industry. And so, I can't say thank you enough, but of course our chair, Dr. Bhatt, we're all in awe of how smoothly you ran this. I just have to maybe voice that for at least what we're seeing here. So, thank you. Thank you all.

Closing Remarks

Dr. Bhatt: Perfect. I get to say my "Thank you," too. This is the one benefit of chairs, I do get to sometimes have a last word, otherwise ordinarily Troy gets it.

First of all, I want to thank the Committee for accepting this call to action because there are a lot of eyes, a lot of opinions, many companies, research labs, friends of ours, and most importantly, patients who are hopeful for the promise of generative AI. And that's why we consider our job here the development of an infrastructure for growth with guardrails.

A huge thank you to Commissioner Califf, the leadership of the FDA. Every one of these individuals in the FDA that you see, they feel strongly about bringing high-quality, safe healthcare to all. And it's not a job for them, it's a calling. So, incredible kudos to them for putting this together.

I do want to remind all the people who are participating by watching that this is not the end, but only the beginning of a process of continuous change. So, how we

prepare these devices, how we prepare the users, how we prepare the environments in which they're going to be used, and then define, monitor, mitigate risk, our framework for surveillance. That's largely happening from all of you who are watching us right now, and we appreciate the work that you do every day in the real world. We're not going to answer all of these questions in a day. Really proud that we are brave enough to disseminate an actionable framework from this, and then iterate on the infrastructure we create with the help of all those people who are watching.

On a personal note, I wanted to thank all of our colleagues on this journey who are watching because it turns out that even though our average age is half-century here, we have parents and family members and friends watching. And sometimes you say, "Why?" And you realize because it really is a landmark moment. It's a paradigm shift in how we're going to deliver healthcare globally, but especially in the United States. And so, for the fact that we understand that to provide scientific rigor, because science has come so far, we need the help of computing power. That's a big step for us to say that. And I'll be here today.

Adjournment

So, thank you all for being here. Thank you to everybody online for your support and tolerating 18 hours of us. And we look forward to more comments up until January 21st. And with that, we will close and adjourn this meeting, the inaugural meeting of the FDA's Digital Health Advisory Committee. Thank you.