UNITED STATES OF AMERICA

FOOD AND DRUG ADMINISTRATION

+++

CENTER FOR DEVICES & RADIOLOGICAL HEALTH

+++

DIGITAL HEALTH ADVISORY COMMITTEE

DAY 1 CONFERENCE

+++

TOTAL PRODUCT LIFECYCLE CONSIDERATIONS FOR GENERATIVE AI-ENABLED DEVICES

+++

November 20, 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	New York, NY

		Chief Diversity Officer Columbia University	
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 Medicine UMass Chan Medical School	Worcester, MA
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 Fulmar, 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: Alright, I would like to call this meeting of the FDA's Digital Health Advisory Committee to order on November 20, 2024. I'm Dr. Ami Bhatt, I'm Chairperson of this Committee. It's an honor to be here. I'm a Medicine and Pediatrics Trained Cardiologist, practiced at the Mass General Hospital starting in 2009, ran outpatient cardiology, telecardiology, and most recently, three years ago, moved to be the Chief Innovation Officer at the American College of Cardiology, our global nonprofit. Again, it's an honor to be here. Important things about today. This is the first Advisory Committee to address digital health technologies and I think it's important to recognize that this Advisory Committee is the most formal and public way that the FDA can receive advice from the public on scientific matters.

Our Digital Health Advisory Committee has members with expertise in various fields of digital health, including machine learning, digital therapeutics, remote patient monitoring, software health data, user experience, real-world evidence, patient-generated health data, interoperability, personalized medicine and genetics, cybersecurity, implementation, and patient experience, along with a lot more. Just looking at the room, including our guests today, it's a really diverse group, and I want to congratulate the FDA on choosing such a diverse group to come together, because we recognize that we need to hear different opinions to make really good decisions.

I'll 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, and for today's agenda, the Committee will discuss and provide 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.

Before we begin, I'd like to set a few ground rules. We have a really busy agenda today, so we'll need to move through it efficiently and with purpose. I ask all presenters to stick to their allotted time. You will see a green, yellow, red button. Everybody else will see that too, and so that will help us make sure we stay on schedule. If any panelist would like to ask a question, please raise your hand or you can put up your card and I will address your question as we move through each session. To ensure smooth communication, please avoid speaking over one another, as this meeting is being transcribed for the official record. Finally, please remember to mute your microphone when you're finished speaking.

I would like to ask our distinguished Committee members and FDA experts to introduce themselves. I'll start by talking to Dr. Diamond here, and Dr. Diamond, maybe after you introduce yourself, we will just go ahead and go down the line and that might save us some time.

Dr. Diamond: Perfect. I'm Dr. Matthew Diamond. I serve as the Chief Medical Officer here at FDA's Digital Health Center of Excellence. Thanks.

Dr. Fulmer: Hi, I'm Sonja Fulmer. I'm the Deputy Director for the Digital Health Center of Excellence.

Mr. Tazbaz: Good morning, everyone. I'm Troy Tazbaz, the Director for Digital Health Center of Excellence at the FDA.

Dr. Radman: Good morning. Thomas Radman. I'm at the National Institute of Health and National Center for Advancing Translational Sciences.

Dr. Khubchandani: Good morning, everyone. Jagdish Khubchandani. Professor of Public Health, New Mexico State University.

Dr. Soni: Hello, everyone. Apurv Soni. Professor and Director of Program in Digital Medicine from UMass Chan Medical School.

Dr. Rariy: Hello, good morning. Chevon Rariy. I'm a Physician and Technology Executive with expertise in leading multidisciplinary teams really at the intersection of care delivery, data analytics, and digital health. I am excited to be here today and look forward to the discussion. Thank you.

Dr. Maddox: Good morning. Tom Maddox. I'm a Cardiologist and Professor at Washington University School of Medicine in St. Louis and I lead the Healthcare Innovation Lab at the school and its partner health system BJC HealthCare.

Mr. Swink: James Swink. I'm the Designated Federal Officer and team lead for the CDRH's Advisory Committee.

Dr. Jackson: Good morning. Jessica Jackson. I'm a Licensed Psychologist and I work with startups leveraging GenAI, digital therapeutics, and virtual reality for health innovation.

Dr. Shah: Morning. I'm Pratik Shah. I'm a Faculty Member and Professor at the University of California and my expertise is generative AI technologies and their deployment. Thank you.

Mr. Posnack: Good morning. Steve Posnack from the HHS's Assistant Secretary for Technology Policy. I serve as Principal Deputy.

Dr. Kukafka: Good morning. Rita Kukafka. I'm a Professor of Biomedical Informatics at Columbia University.

Dr. Elkin: Hello, I'm Peter Elkin. I'm a Distinguished Professor and Chair of the Department of Biomedical Informatics, the University of Buffalo, and also involved in doing research in AI in general, and I trained at MGH, so there you go. Thank you very much.

Dr. Botsis: Taxiarchis Botsis, Associate Professor of Oncology Medicine at Johns Hopkins University School of Medicine, also Director of the Biomedical Informatics Score at the Cancer Center.

Dr. Stanley: Hi, everyone. I'm Laura Stanley. I'm a Professor in Computer Science at Montana State University, Bozeman, and happy to be here.

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

Ms. Miller: Good morning. Diana Miller. Senior Director of Data Science at Medtronic. I'm here as an Industry Representative. Happy to be here and meet everybody.

Conflict of Interest Statement

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

Mr. Swink: Thank you. Good morning. I will now read the FDA Conflict of Interest Disclosure Statement. So, 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 of the status of this Committee's compliance with federal ethics and conflict of interest laws are covered by, but not limited to, those found at 18 U.S.C., Section 208, and 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., Section 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 the discussions of today's meeting, members and consultants of this Committee who are 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., Section 208, their employers. These interests may include investments; consulting; expert witness testimony; contracts, grants, CRADAs; teaching, speaking, writing; patents and 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 premarket performance evaluation, risk management, and postmarket performance monitoring for generative AI-enabled devices.

Based on the agenda for today's meeting and all financial interests reported by the Committee, the members and consultants have-- No conflict-of-interest waivers have been issued in accordance with 18 U.S.C., Section 208.

Ms. Diana Miller is serving as an Industry Representative for generative AI/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 Dr. Danielle Bitterman, who is an invited guest speaker with us today, has acknowledged possible

interests through her employer, Brigham and Women's Physician Organization, and in the form of federal research grants, and as a scientific advisor with an unaffected firm. Dr. Bitterman also acknowledged that she is an Associate Editor of the Radiation Oncology for HemOnc.org and she receives no financial compensation for that activity. Dr. Faisal Mahmood, who is also an invited guest speaker with us today, has acknowledged a possible interest through his employer, Massachusetts General Brigham, and has acknowledged interests with an unaffected firm in the form of a stock and as a scientific advisor. Dr. Pranav Rajpurkar, another invited guest speaker with us, has acknowledged interests with an unaffected firm as a Co-Founder of the firm. Drs. Keith Dreyer and Gabriella Waters are other invited guest speakers with us today and have reported no interests 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 the 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 relationships that they may have with any firms at issue. A copy of this statement will be available for review at the registration table.

Before I begin, I would like to make a few general announcements. We have transcripts of today's meeting; will be available from ACSI. Their phone number is 202-599-8456. Information on purchasing videos of today's meeting and handouts for today's presentations are available at the registration table outside of the meeting room. The FDA press contact for today's meeting is James McKinney. I think he'll be-- Here he is, over here on this side. All written comments received were provided to the Panel and the FDA review team for their review prior to today's meeting. There is an active

docket where members of the public can post written comments. The link can be found on the FDA website and the registration table, and that will be open for another-- I think until January 30, so you have time to submit more comments. If you are presenting in the Open Public Hearing Session and have not previously provided an electronic copy of your slide presentations, please arrange to do so with Mr. Artair Mallett at the registration table. And in order to keep the transcriptionist-- Please identify yourself before you speak each and every time. Please, silence your cell phones and other electronic devices. Also, we made handouts for both days. If you took handouts, please bring them tomorrow as we have limited copies. Thank you.

Welcome from FDA Commissioner

Dr. Bhatt: Thanks so much, James, and thanks for all the work you did to get us all here today. So, we will proceed with welcoming remarks. We are honored to have the FDA's Commissioner, Dr. Robert Califf here with us. Dr. Califf.

Dr. Califf: Hey, it works. Alright. Good morning, everybody, and welcome to this inaugural Digital Health Advisory Committee meeting. I am pleased to speak with you today and I want to thank all the experts who are serving as members of the Advisory Committee; our CDRH team for their thoughtful planning of the meeting; and our speakers for sharing their perspectives. I also want to thank the many product developers, representatives of industry, academia, and consumer organizations, as well as the patient advocates and others who are participating today. Advisory Committees are a unique mechanism that help support the FDA's mission. They bring together leading experts who can advise the FDA about a particular field, in this case digital health technologies, giving us access to the latest thinking on broad issues and how to navigate them. These forums also encourage the public, including patients, healthcare providers, industry, and other interested people to share their views.

The Digital Health Advisory Committee will provide advice and recommendations on understanding the benefits, risks, and clinical outcomes associated with digital health technologies. We've established this Committee because we see great potential for digital health technologies to help address critical healthcare issues that we face today, and we need these technologies to be developed, deployed, and used responsibly in the best interest of patients and consumers. I've talked before about the current catastrophic decline in life expectancy in the United States, leaving us years behind our economic peers in life expectancy and ability of our people to function. We are even falling behind many middle-income countries like Thailand and Costa Rica, and just as worrisome and a critical issue for AI is that the negative trajectory is largely driven by disparities that are a function of race, ethnicity, education, and wealth, as well as where someone lives, particularly rural and underserved urban community status.

Artificial intelligence is changing how we think about health and healthcare, and it's one of the most exciting and promising areas of science because it's built to transcend boundaries. While we often describe the promise of AI in terms of helping us facilitate speedier delivery of new treatments, I believe the larger role for AI is to improve the efficient and coordinated delivery of care to patients in all facets of healthcare settings, including the operating room, the clinic, and even the home. Digital tools and AI have the potential to help improve quality of life and life expectancy, including prevention of illness in healthy people for all people living in the United States. To this end, I hope you'll consider CDRH's "Home as a Healthcare Hub" project during this meeting. The FDA regulates 20% of the U.S. economy and must continuously evolve to oversee the safe and effective use of AI across regulated industry, ensuring compliance while fostering innovation.

For us to be most effective and do our jobs as protectors of public health, it's essential that we embrace these groundbreaking technologies, not only to keep pace with the industries we regulate, but also to use regulatory channels and oversight to improve the chance that they will be applied effectively, consistently, and fairly. I can't get Ed Yong's phrase from his article in the Atlantic out of my mind. Technological advances drift into society's penthouses. Pandemics—we can substitute diseases for pandemics—seep into its cracks. Unless we have intentionality to reach technologically disadvantaged people, we'll continue to even-- Or even accelerate the current trajectory to further widen a huge gap in health and life expectancy between the wealthy and highly educated people, mostly living in urban areas, and people with less wealth in education increasingly dominated by the rural population. The FDA has been working for years to anticipate and prepare for challenges of AI and to harness the potential to enable major advances in the development of more effective and risky-- And less risky medical products.

The FDA's first approval of an AI-enabled medical device took place almost 30 years ago in 1995. Since then, we've received approximately a thousand submissions for AI-enabled medical devices and more than 300 submissions for drugs and biological products with AI components. But I have to say, I'm also told on the drug and biologic side almost every application now has AI somewhere in the development process. These submissions have included drug discovery and repurposing, enhancing clinical trial design elements, dose optimization, endpoint and biomarker assessment, and postmarket surveillance. They also cover a growing diversity of medical devices that leverage AI to improve clinical workflows in patient experiences or outcomes in addition to sophisticated prediction algorithms. For example, the Agency recently reviewed and granted a new AI algorithm that supports earlier detection and

management of hypertrophic cardiomyopathy, something dear to my heart as a cardiologist. This granting created a new regulation for cardiovascular machine learning-based notification software and will help provide quicker identification of possible suspected HCM cases allowing for earlier patient treatment, hopefully before the first episode of sudden death.

We recently published a paper on how FDA's CBER, CDER, CDRH, and Office of Clinical Policy are working together on AI and medical products. The paper outlines and reaffirms our commitment to promoting the responsible and ethical development, deployment, use, and maintenance of safe and effective medical products that incorporate or are developed with AI.

Two final notes for your consideration. First, as I talk with clinicians across the country, I'm hearing increasing concern that the criteria for adopting an AI are almost purely financial. I hope this Committee will help us figure out how to balance patient outcome benefit with financial return. Financial return is relatively easy to measure these days, and AI is a supercharged financial tool. Clinical outcome measurement is hard. Unless you take this issue very seriously and form alliances of those concerned about improving health outcomes, this technology will improve profits at the cost of continued deterioration in our overall health status. At FDA, our statutes prevent us from considering cost in individual product decisions, so we're all in on this at the FDA, but we need your help on this issue because the tide is going the other direction.

Second, I'm convinced that we need a very different information ecosystem to ensure that local recurrent validation of AIs occurs as a continuous process in the midst of our ongoing healthcare system. This is the ultimate in real-world data and evidence, but I challenge you to envision how such a system would work and to focus on the specific examples you're discussing today. I have looked far and wide; I do not believe

there's a single health system in the United States that's capable of validating an AI algorithm that's put into place in a clinical care system. Beautiful operating characteristics of an AI in the premarket phase in no way should assure us of the performance in real life. Imagine that we developed apps for driving in a hypothetical space and didn't incorporate the evolving experience of people who are actually driving.

So, this is well beyond FDA and it should be. We don't have inspectors, for example, on every farm every day. But while not perfect, our farms adhere to the 10 FSMA rules producing safe food for the country. We need something similar for AI and healthcare. You probably don't think of yourselves as farmers, but I believe that you really should in this regard.

I really do wish, and you could probably tell this, that I could stay for the two days to hear the robust conversations in this meeting. This Advisory Committee is a valuable resource and a forum for contributing to how we think about this technology to ensure that devices are safe and effective. We're seeking your input to help inform the FDA's thinking as we work to ensure that people living in the United States benefit, not just financially, but also with regard to their health outcomes, from AI's promise to improve healthcare.

Thanks for being here today. I hate to talk and run, but I will tune in. This is going to be on YouTube, right? Alright. Well, I take this in while I'm trying to sleep at night. Thank you.

Opening Remarks

Dr. Bhatt: Dr. Califf, thank you so much for your remarks. Your insights are perfect, as we really think about this critically important topic. We'll proceed with opening remarks from Dr. Michelle Tarver, Director of CDRH at the FDA. Dr. Tarver, thank you for being here with us this morning.

Dr. Tarver: Well, thank you for having me. Good morning, everyone. I want to thank Dr. Califf for his remarks and I want to welcome you all; those of you on the Panel, those of you in the room, those of you tuning in online. Today's an exciting day. It's our first Digital Health Advisory Committee meeting. This is a milestone and it really does come in the context of how we're seeing healthcare, medical devices, and patients' expectations evolve over time. You've probably heard us say at the FDA "Patients are at the heart of what we do," and we really mean it. The Center for Devices and Radiological Health's vision statement begins with a commitment to the American people that patients in the U.S. have access to high quality, safe and effective medical devices of public health importance first in the world. And we don't say first in the world out of a spirit of competitiveness, but truly out of a pledge to timely access.

When we're talking about public health, every patient counts and every day matters. And it's by working collaboratively to foster innovation and to facilitate timely access that we can have true impacts on this nation's health. Now, you heard Dr. Califf talk a lot about the flood of medical devices that we have coming into our doors, and many of them are AI-enabled, and AI holds the possibility of transforming how we prevent, diagnose, and treat health conditions. And what's so exciting about it is that it continues to change and evolve each day, offering new opportunities for us to deploy it in the care of patients, but only if it's done responsibly and we can rely on it to be safe and effective. And it's going to be important not only in a traditional clinical setting, but also outside of the clinical setting, like the home and communities, closer, where they don't have clinics; where they don't have hospitals.

In fact, if you look at the CDC reports, currently there are 129 million people in this country that are living with chronic diseases, and 40% of them have two or more chronic diseases. That accounts for over 90% of the 4.1 trillion dollars we spend on

healthcare expenses. Huge population; huge cost. And that's in the setting where we have consolidation of health services, where we see contraction of providers and facilities. How are people going to get the care that they need? Well, AI-enabled medical devices offers a promise to extend into those communities and offer options for them to get care. When we're thinking about the burden of disease and who is most bearing that disease; that's small-town America. Those are rural communities. There are people who are of older age, racial and ethnic minorities, those with fewer resources and who live the furthest from healthcare facilities. So, technologies have to come meet them where they are. You heard Dr. Califf allude to our initiative on "Home as a Healthcare Hub", and it is really grounded on that principle of how can we democratize care in a way that everyone has access to it; that everyone can benefit from the healthcare-- Good healthcare outcomes that all of us would like to experience.

One of the things that I'm really excited about today's conversation is this is really a collective conversation on how do we move things forward in an equitable and ethical way. And that requires including everyone at the table. That means patients, providers, industry, regulators, developers to have a collective conversation. Well, today we're going to talk about generative AI and the total product lifecycle of these AI-enabled technologies, particularly looking at the benefits, the risks, the outcomes, and what are the considerations as we look at them both as regulators and ultimately how you all will be looking at them as care providers and patients.

One of the things that I want to definitely highlight is how important it is for us to be having the conversation now and not having it five years from now, when we're trying to figure out how to manage it in retrospect. So, being proactive, being deliberate, and being intentional in this conversation really will help us make informed decisions.

I'm really looking forward to the next two days. I'm really looking forward to the conversation, and I'm not going to keep you listening to me for very long, so I'm going to turn it over to Troy Tazbaz, who is the Director of the Digital Health Center of Excellence to give some remarks. So, thank you all very much for being here. Thank you all for your insights today and I'm looking forward to the conversation.

FDA Perspective - GenAI in Medical Devices

Dr. Bhatt: Thank you, Dr. Tarver. As Mr. Tazbaz starts, just for the Panel to know, we are now going to have the first five presentations of our agenda and we will then be able to ask questions after we have the next five Panel presentations. Mr. Tazbaz, thank you.

Mr. Tazbaz: Thank you. Good morning, everyone. I first want to extend my heartfelt gratitude for all of you, for being here and dedicating your time, your knowledge and expertise to these very, very critical discussions. I also want to thank the FDA team for working tirelessly to making this event happen. It was difficult, but it is going to be very well worth it, and we're very excited about the discussions that are going to be taking place today.

The Digital Health Advisory Committee has attracted significant interest from the entire healthcare community. We believe this is a reflection of the growing role digital health technologies will play in advancing patient care and overall health of our entire population. As the FDA's first Advisory Committee focused on crosscutting digital health issues, we hope DHAC, as we've been calling it, will guide us on matters that affect the wide spectrum of medical device products.

Today we are here to address one of the most transformative developments in healthcare: generative AI, or as we've been referring to as GenAI. We hope to explore and identify unique regulatory considerations necessary to ensure that safety and

effectiveness of these rapidly evolving technologies are maintained throughout their lifecycle. Generative AI brings incredible opportunities to enhance patient care but also introduces new regulatory challenges and oversight-- And operational oversight. As it's been highly covered, generative AI has the potential to reshape healthcare by generating contextually relevant outputs from vast amounts of data. However, unlike traditional AI models that predict or classify data, GenAI produces new outputs which can introduce additional layers of regulatory complexity for us. This is especially relevant when these outputs could directly impact the safety and effectiveness of medical devices.

To guide us throughout the discussions, we would like to take a total product lifecycle approach, as Michelle mentioned. The lifecycle perspective allows us to consider safety and performance from initial design and development all the way through the postmarket monitoring. Throughout the next two days, we will dive into critical questions surrounding generative AI, including premarket evaluation, performance metrics, and effective postmarket strategies. Additionally, we will discuss the complexities of monitoring generative AI models across multiple sites, as we need to ensure that they perform consistently while managing potential regional biases and data variations.

Our agenda is designed to give structure to these important discussions, beginning with an examination of information needed to evaluate generative AI-enabled devices, followed by in-depth sessions on risk management, and postmarket performance monitoring strategies, which is a very important part of the equation of when these things are deployed into real healthcare settings.

As we move forward, I'm confident that the insights and ideas shared here will shape the FDA's approach to balancing innovation with safety and ensuring that the patients and healthcare professionals can trust these powerful new tools.

I would like to say thank you once again for your participation and engagement. I look forward to the rich discussions ahead and the actionable steps we will identify here together. In a moment, you will also hear from Dr. Aldo Badano on some regulatory science challenges that FDA faces for evaluating generative AI applications and medical devices, and tomorrow FDA's Jessica Paulsen will provide an overview of approaches for managing changes for AI-enabled devices. Thank you, everyone.

Sub-topic: Premarket Performance Evaluation

FDA Perspective – Regulatory Science Challenges for the Evaluation of Generative AI Applications in Medical Devices

Dr. Bhatt: Thank you, Mr. Tazbaz. We will now proceed with Dr. Aldo Badano's presentation. Dr. Badano, as mentioned, is Director of the Division of Imaging, Diagnostics and Software Reliability in the Offices of Science and Engineering Laboratories at CDRH-FDA. Dr. Badano.

Dr. Badano: Good morning. Thank you for the introduction and it's an honor to speak to the distinguished members of the Panel in the audience.

Let me start by saying a few words about our office. OSEL in CDRH, Office of Science and Engineering Labs, is an ISO 9,001-certified research organization dedicated to promoting innovation for the development of new life science medical devices. Also, it is organized into about 20 program areas and AI/ML is one of the largest programs that we have. Also, outputs are regulatory science tools, which are innovative open-source tools for assessing the safety and effectiveness of emerging technology. Innovators can use these tools voluntarily into all stages of device development. Regulatory science tools have been cited in over 1,000 applications-- Premarket applications to the Agency, over a hundred product codes. If you would like to learn

more, there's some information about how to reach our regulatory science tools catalog with a website and the QR code.

GenAI-enabled devices in healthcare have a wide range of uses. Here's a table taken from the literature; an article by Meskó and Topol, which outlines some of the applications and how the community is thinking about GenAI in the medical space. Some of these products are clearly nonmedical devices, but there's a lot of excitement in terms of the potential performance and productivity gains. However, the same characteristics that make these devices and products have so much potential are the characteristics that might pose challenges for certain uses.

So, what are the regulatory science challenges that this technology brings about? Here's a preliminary list. I'm not going to go into detail in all of these; we don't have the time. But many of these challenges might exist in other technologies, but are exacerbated in AI applications, particularly in GenAI. So, difficulty in defining scope of the products intended use; some of the parts of the model like foundation models not being in the provenance of the manufacturers of the device; oversight of adaptive systems; and hallucinations, which are a new kind of error that has been introduced in recent years by deep learning and other technologies that bring about the new trade off between stability and accuracy due to concerns about reliability and trustworthiness.

There's also issues about data needs, including diversity; monitoring performance in the real world, including bias; transparency to users. So, let's narrow it down a little bit with one example use case. In this use case, a GenAI-enabled radiology device can produce a report when prompted with a medical image, as you see in the left, in the input side. The reference for this device might be a clinician-generated report. So, what strategies do we have at hand to evaluate these new technologies?

In recent years or months, I would say, three distinct classes of evaluation have emerged from the scientific literature. Benchmarking with the standardized reference datasets; expert evaluation, including a holistic evaluation of the AI-human interaction; and model-based evaluations, which include automated testing.

In the next slides, I'll describe a little bit the pros and cons of each one of these approaches. Benchmarking. Benchmarking is the evaluation of models on external testing datasets with predetermined evaluation metrics. This approach is practical, allows for head-to-head comparisons of models on same data and with same metrics, and can assess models at a large scale. However, benchmarks are limited in models and in tasks or applications, and there is this risk of that the models can train to the test, which is an overfitting technique for optimizing performance on a given dataset at the expense of performance for other data distributions, for instance, like the ones you see in a different clinical setting. On the right, you can see one of the many leaderboards available in open-source format.

Expert evaluation. Expert evaluation refers to using experts to establish the reference standard for a given task. For example, experts may determine if an error from GenAI will cause significant patient harm. Note that in the chart, experts can identify errors that might pass tests of a benchmark concern only with tech similarity metrics. For instance, right versus left, which is location of disease; fusion versus effusion. Expert evaluation is adaptable to a wide range of tasks with direct clinical relevancy. However, expert evaluation is resource-intensive and burdensome, and must consider the subjective nature of the readers and the highly variable effects.

And finally, model-based evaluation. This is about evaluating models with other models using a model-based approach that must have human oversight. Let me bring up the chart. The advantages are clear. This augments the human evaluation and is scalable.

However, the disadvantages include the fact that the validation needs to be robust and there is a potential risk of inter-model leakage, which means models in the device and in the evaluator can be correlated, therefore generating risks of the results. In our CDRH labs, we're actively investigating a model-based evaluator for factual accuracy of a GenAI-generated radiologic impression. A series of binary questions are used to determine the rates of different types of errors in the generated report. We're also working on performance metrics and associated statistical analysis plans. I hope that you'll stay tuned for updates in the coming months.

Let me summarize. For some GenAI, known evaluation strategies still may apply. However, new evaluation methodologies and new performance metrics may need to be developed for some other types of GenAI. Overall device performance evaluation requirements are governed by the intended use and the associated risk for the device. And finally, particularly in the case of GenAI-enabled devices, evidence may include pre and postmarket elements as the technology is meant to update often based on the availability of models that are trained on more and more data and the emergence of novel architectures and GenAI innovations.

In summary, rigorous evaluation of GenAI-enabled devices is necessary as a rapidly emerging technology developing a least burdensome approach to ensure safety and effectiveness will require partnerships like the one we're seeing here in this room and commitments to advancing regulatory science. Thank you for your attention.

Stakeholder Perspective – The Usage of Generative AI in Digital Pathology and Potential Challenges for Evaluation

Dr. Bhatt: Thank you, Dr. Badano. Next, we're going to hear a stakeholder perspective from Dr. Faisal Mahmood. Dr. Mahmood is Associate Professor at Harvard University Brigham and Women's Hospital. And just a reminder to the Panel that when

we are done with these, we'll take a short break and then we can ask questions of the panelists; that includes Mr. Tazbaz and Dr. Badano, so to save your questions for them as well as our next three speakers.

Dr. Mahmood: Okay, thank you for having me today. So, I'll be talking a little bit about some of the work that my group has done over the years in computational pathology and how it has led into generative AI and what we are currently focused on. Just a quick outline for the talk. I'll be talking about some of our older work and how it feeds into the generative AI-related work that we have done more recently.

So, the problem that we are trying to target is to analyze pathology images. Pathology images are hierarchical, they're large. They're very large gigapixel images that result from the digitization of glass slides. And the goal of our field is really to go from these images to everything in the red box here. This involves early diagnosis, prognosis, prediction of response to treatment, outcome prediction, patient stratification, and some form of biomarker discovery.

So, I'll start by talking about some of the "historical work," even though it's quite recent. So, this is some work that we did back in 2021. This allowed us to use these large whole-slide images and slide-level labels that existed in pathology reports while still being data efficient, so without requiring any pixel level and annotations. And this was published back in 2021 and we made the code and the models, everything, publicly available for the community to use. And these tools have been extensively used and have basically been applied to every major disease organ system. And we applied it to a number of different applications, including for cancers of unknown primary, where we tried to find the origin of the tumor. It's a very common and difficult problem in pathology, and we posed this as being an 18-class supervised classification problem using host-led images and labels that were taken from a pathology report.

Overall, we found that this approach works quite well. Internal and external test sets; we did a lot of evaluation assessing the variability. And then more recently we have tried to combine histology images with molecular data to see if we can improve outcomes and have shown that this is also possible. We're also interested in integrating histology with molecular information to improve outcomes in general, and have shown in this study from back in 2022 that you can improve patient outcome stratification into distinct risk groups, but then also go back and identify biomarkers that would be relevant to the diagnosis but also would correlate with response and resistance to specific therapeutic agents.

This is another example for endomyocardial biopsy assessments. Our approach has been that we can train these supervised models, but then targeting problems that are difficult to solve in pathology. So hence they would sort of encourage machine learning adoption because these problems are difficult to begin with, like cancer of unknown primary or endomyocardial biopsy assessment where there is large scale intra and intra observer variability.

But the purpose of showing all of these examples is that they all follow this very conventional approach that's commonly used in computational pathology, where we have whole-slide images that are digitized; we patch them because they're very large; we extract relevant features, or in this case, they're exhaustively extracted from a pre-trained encoder, leading to aggregation and eventually a prediction. This pipeline has been innovated on in various components over the years, both by the clinical community as well as in terms of how the data is used and by the machine learning community and how we feature, extract, and aggregate.

But over the years, we have found that the feature extraction component is the most important component, and that's what leads us into foundation models or self-

supervised models. In the examples I showed earlier, we were using a model that was trained just on a ResNet that was trained on real-world images from the ImageNet database and is used for feature extraction. But more recently we've seen that self-supervised models perform significantly better. So, motivating foundation models for pathology. So, foundation models are generic models that are capable of rich feature extraction. The particular use that we have here is that by having richer feature extraction, we can apply this to situations where we just don't have enough data; very, very common in pathology. For example, for rare diseases, clinical trials, and they're ideal for multi-task, multi-tissue kind of problems, and they're not necessarily meant to replace end-to-end trained supervised models. So, this thinking was sort of what led to these two studies. Those from March this year, where we have two foundation models. The first one, UNI, is a self-supervised model that's trained in a self-supervised manner on lots and lots of pathology data to extract rich feature representations from those pathology images. And CONCH is a visual-language model which further enhances this by contrasting with text. And this is in line with how the machine learning community is thinking about this, because they've shown that contrasting with additional modalities leads to better improved feature representation for both images and for text.

So, we collected a large dataset of about a hundred thousand whole-slide images. This came from a number of different hospitals including our own. And it's diverse; it was maximized for diversity. And this is also in line with how the machine learning community has found that diversity of data is much more important than the quantity of data in building these large self-supervised models. And we used the DINOv2 framework from Meta to train these models, and it was assessed against a number of different benchmarks, both conventional and pathology-related, and it was applied to 33 different downstream tasks. The radar plot is not the best way to show

this, but it does capture all the different tasks that this was assessed on, including for whole-slide level classification as well as for region-level or small-scale classification. And in particular for few-shot classification, where we attempt to train these models with very, very few data points, hence assessing its efficacy on situations where you would not have enough data available, for example, for rare diseases, clinical trials, and so forth. We also wanted to see how far-- I'm sorry, the colors on this slide did not show up as they originally were. But we also wanted to see how far can we push this. Can we have a single model that can cater to a really large difficult classification problem of 108 different cancer types in the entire OncoTree classification system? And the goal here is not to reach clinical utility—it would still take some time; more data, potentially, to reach to get there,—but to really assess that how rich the feature representations get if we use this form of self-supervised learning.

And in parallel, we were also working on contrasting images with text. And in this case we use 1.1 million image-caption pairs that come from the PubMed open access database. We could not use all the data in PubMed, only that that was publicly funded and is within the open access database with an appropriate license. But contrasting images with text is a very common approach used by the computer vision community to enhance feature representations for both images and for text. This is a distribution of all the data that was used in this particular case. And we assess this in a number of different scenarios, including for Zero-Shot classification, where the goal is to really assess the feature representation and not really achieve clinical utility, but also for Few-Shot classification, we're showing that fewer data points can lead to better performance and a variety of different disease models, both rare and common.

These models were made publicly available and we essentially did this to comply with NIH guidelines and what the requirements were from the journals where

we were publishing this. But the models were extensively downloaded and have been used all over the world for a host of different applications or for every major organ system all the way up to forensics. So, things that we never originally thought of. We didn't even know that this many people were working in computational pathology.

And they have also been assessed independently by a number of different groups. This is an evaluation from a group in Dresden, where they have shown how these models work against a number of other approaches, self-supervised and otherwise, for a host of different tasks including diagnostic tasks, as well as predicting what's commonly referred to as non-human identifiable features, like predicting molecular markers and other alterations directly from histology images and just prognosticating directly from histology images.

Here's another assessment that was done at Mount Sinai and at the Memorial Sloan Kettering Cancer Center using our foundation models and how well it fares for some of these very difficult pan-cancer predictive tasks. Since these models came out, we have further expanded them to a whole-slide level. So, the models I showed so far work at original level extracting rich feature representations from regions. We have since expanded this to extracting a single feature vector that corresponds to the whole slide, and we've done that on an image level. And then it's further contrasted with generated captions that come from our generative model that I'll talk about in a second, but also from the pathology report. So, contrasting both with a generative caption and a pathology report. And we find that this form of feature extraction, where we have a single feature vector for the whole slide, essentially enables a lot of different things that we can do downstream for incredibly difficult tasks and for common clinical tasks. For common clinical tasks, we can get clinical-grade performance and for more difficult tasks, we're further trying to improve the model to get there.

But I think what most people get very excited by is that-- Its ability to perform on Few-Shot examples in situations where diagnosis is incredibly difficult. Average diagnosis for rare disease is between six to eight months, and it really is a challenging problem where we can use these rich feature representations to find similar cases, but also train with very, very few cases. And then of course, because it has a text head, we're able to generate these reports. We compare it with the pathologist-given reports, and we've done quite some analysis on how good the model is for report synthesis, both at the level of the slide for the patient, what happens when we include electronic medical records, and so forth.

So, the story so far is that we started with lots of supervised models. We had self-supervised models or foundation models that relied on images themselves, and then we contrasted those images with text to improve the image feature representation. So, there are a lot of other modalities that we have in pathology and in healthcare in general that can be used to contrast data to further improve feature representation for each one of these individual modalities. So, we contrast-- In this example—this is an article from earlier this year,—we contrast pathology images with immunohistochemistry, and by contrasting we further improve the feature representations with just a single model that can extract rich features. And then it's applied downstream to a variety of different tasks, including IHC quantifications with the first time showing that you don't need to have pixel-level annotations for this very difficult task of quantifying IHCs. And it could be just this model trained on lots of IHC data that can cater to all of immunohistochemistry scoring as well as downstream applying it to survival and so forth.

And if this is true that we can contrast with IHC, an obvious next step would be that can we also contrast with transcriptomics? So, in this smaller example that was

published earlier this year in CVPR, we showed that we can do this and it can substantially improve Few-Shot performance for detecting rare diseases. But more recently we have expanded this to a very, very large-scale study where we are using all the molecular data from the Brigham and MGH that's been collected over the years and correlated with histology images. And by contrasting with our NGS data, with transcriptomic data, and with histology data, we were further able to improve feature representations. And these are a variety of different downstream tasks that it was applied to, but just to draw your attention, it was able to improve performance for these incredibly difficult tasks for predicting treatment response, so often for which we have very, very few samples available because at a tertiary medical center, like ours, it's often difficult to get pretreatment biopsies, because a lot of patients get referred rather than the hospital being their first instance. But given the data is so small and the downstream treatment efficacy is still low for a variety of different therapeutics, improving performance by using image-based biomarkers is of a lot of interest to drug companies and other stakeholders. So, this is quite exciting for us to see that for the first time we can just boost performance by using these contrasted foundation models.

So, what do we do after we have all these foundation models? We can of course train lots and lots of supervised models by using these self-supervised rich features. But what was most interesting to us is that we can also use them in a generative setting where we can have a singular model that can cater to all of human pathology. And this was the biggest undertaking that my group had done. So, our entire group was involved in collecting the data for this, because the data required is not collected essentially; essentially clinically. So, our goal is to build a singular, multimodal large language model that can cater to all of human pathology. And why we believe this is possible is because there are increasing number of large language models and multimodal large

language models that cater to very, very large amount of information. Like open AIs trying to build a single model that could cater to the world's information, and human pathology knowledge is much, much more constrained and specific. So, if the first is true, we should be able to build a model that caters to just all of human pathology. And what we need to do is that we need rich feature representation that can come from foundation models or self-supervised models; we need a large instruction dataset of images, text, and corresponding responses; and we also need robust evaluation.

The challenge we face is that we have these hierarchical large pathology images, but the annotations we have corresponding these images are at the level of the slide and not at region level, whereas an effective chatbot for pathology or a copilot for pathology or a singular model that can cater to all of pathology will have to operate at all of these levels. So, the single biggest challenge is that fine grained understanding of pathology regions and cellular level information leads to slide-level diagnosis. But that fine grained information does not reside in pathology reports. So, this required a large amount of manual data collection, which we had from multiple collaborators across six, seven different institutions. And the data collected also needs to be incredibly diverse and should cater to all known entities within, of course, neoplastic, infectious, inflammatory or otherwise, that would be added.

So, in this initial version that was published in June of this year, we had about 999,000 question and answer turns, corresponding images, questions and answers that were used to train this model. So, this relied-- The image features came from the foundation models that we had developed earlier, and then it was fine-tuned on this large 1-million-instruction dataset. And then this led to PathChat, which is a generative AI tool that can cater to all of human pathology. It catered to a smaller subset back when the article was published, but now we have expanded this to about 13 million

question and answer turns and have really extended the data substantially with a lot of institutions volunteering data just by virtue of the fact that this made a bit of noise in the community. But you can ask basic questions like, “What am I looking at here?” So, it could be a very good training tool. But what we are most excited about is that it can basically work at a whole-slide level where it can generate the report and by the time a pathologist is looking at it, they would have the entire trajectory of the diagnosis worked out and that report already generated.

Another thing that we are very excited about is that it can ingest multiple images within the same context, which means that if-- by looking at an image, a pathologist decides that additional tests would need to be ordered. Those tests could automatically be ordered. And once those images from, for example, immunohistochemistry are received, they can be ingested into the same context and our report would already be generated. So, in this example here, the report is being generated. Of course, a big utility. We had a lot of interest from a variety of different foundations who would like to use this tool in lower source settings because there are very limited pathologists in resource-limited settings, and it’s the norm to send or ship out slides for a diagnosis to be made in resource-rich settings. However, that’s prone to a lot of delays and it’s been shown that such cases, depending on where they’re sent and how much data-- How much tissue material is available, also have a lower accurate diagnosis rate.

So, in this case, the user is just taking an image with a cell phone coupled to a microscope. It’s a very common approach used in lower-source settings, and people take these images and send them to their colleagues in resource-rich settings and trying to see how well this worked. We did a lot of domain adaptation analysis, trying to see that these models that are trained largely on whole-slide scanners from a variety of different vendors, how they would adapt to image generation in this form and how the

resolution would vary. And then went back to the drawing board and improved our training pipelines to cater to some of those variabilities that would improve domain adaptation.

Another example, and something that a lot of people are very excited by is our ability to use PathChat by just a microscope coupled to a camera. So only about 4% of pathology diagnosis in the U.S. is digital, or 90% of the pathologists still use a glass light under a microscope. Most microscopes in the U.S. have a camera attached to them. So, in this case, the user is asking that-- This is a lymph node mass, and what's the likely diagnosis? And the chatbot says that, well, it's likely a melanoma. And the user can ask that what IHCs or additional ancillary tests could be used to confirm this. And we have a number of different possibilities. And once those tests are in, you can take an image of the IHC and it would ingest it within the same context. And this capability is possible because during the training, we mapped out trajectory for over a hundred thousand cases. Of course, we had lots of one-shot training with lots of images, questions and answers. But also, we mapped out entire trajectory of cases where we start with an image, additional tests are ordered, and basically the entire diagnostic trajectory of the entire case, and use that as the training data. And we are further assessing this capability on a large dataset that's collected across multiple institutions.

The most important component in building this-- Of course, building this took about three years. But in building this, the most important component for us is the assessment and evaluation, and how the evaluation needs to evolve over time as this chatbot essentially gets better. So initially, this was largely based on-- The qualitative analysis is based on multiple-choice questions. We basically took inspiration from two different domains; inspiration from how generative machine learning is evaluated for conventional, like non-healthcare related machine learning, and how large models that

cater to multiple diseases have been evaluated on the healthcare side, like NGS essays and so forth, and brought the two together.

So, in this particular case, we're looking at lots of multiple-choice questions and how the model fares against other generative models that are not specific to healthcare or are specific to healthcare, but not specific to pathology. But I think because this is generative AI and it gives open-ended responses, it's very important for us to assess how well those responses are given and recorded. So, in this particular case, we had a Panel of seven Board-certified pathologists who assessed the entire trajectory of conversation and each response generated by the chatbot and compared it against other generative models. So, we collected both quantitative responses, whether the response was correct or not—the binary response,— but also a more qualitative response that how well does this-- Does in comparison to other models.

If we go and look at how other large language models are often assessed, it's often done in a competitive setting where multiple models are assessed against often a human. And we found that the models do quite well for their intended use, for diagnosis, and for describing morphologic descriptions, but don't do very well correlating back to clinical or just marginal decrease in performance, or for example, ChatGPT or GPT4V or GPT-4.0. And the reason could be-- We don't know, of course, because we don't know what GPT-4.0 is trained on, but likely OpenAI is training it on all of medical literature and correlating back to clinical might be much more easier in their case.

We're improving the capabilities in terms of the ancillary testing, and this is what was done for the article. But we have substantially expanded this to about 11,000 cases versus 260 earlier and have also expanded the number of pathologists who are evaluating this as we continuously improve this model.

So, now that we have the generative AI model, a next step is obviously to look towards agents. So, we started with supervised models and then we built self-supervised models for literature extraction, and we used those self-supervised models for a generative AI tool. And the next step is to see if we can build an agent so AI agents can essentially do things for you. In our case, we do a lot of biomedical data analysis. We want to see that can we replicate what essentially we are doing every day and could be continuous discovery agents. So, I'd just like to show this one example that we have where the user is saying, "Here's my dataset and I have a cord of responders versus non-responders. Can you go train a model that would predict what the-- Can you go train a model to predict responders versus non-responders?" So, the model would essentially make a plan. So, this plan is essentially coming from the generative machine learning component where you would segment the tissue, extract the feature as train the models and the mechanism of evaluating this. But then it can make use of existing code and patch it together with some additional code that might need to be written. So, in this case uses PathChat for morphologic description, it uses GPT-4.0 for writing the additional code that might be needed, and a host of different libraries that are out there look at the-- To essentially conduct all of these independent tasks. So, once the model is trained, we can ask additional questions. This essentially enables discovery from people who would need a clinical or a computational collaborator to run some of these experiments. And then the other benefit of this is that it can be crawling through large amounts of data, essentially doing what we do with lots and lots of pathology data manually in pathology departments throughout the history of pathology. The way new morphologic diseases are identified-- Morphologic entities that correlate with disease or outcome are identified is that someone goes in, looks at lots and lots of different cases and says "This similar morphology across cases correlates with outcome." In this case,

it's doing this essentially on its own and is using PathChat to generate a report. So, this is quite exciting and we have this being tested at everything we are scanning at the Brigham and Women's Hospital.

Towards the end, I just want to touch upon bias and fairness in computational pathology. So, this is a topic that we are very interested in because we want to make sure that the models that we develop have the broadest reach and as we go from more specific models to more generalized models, we want to make sure that their adaptability across-- Is not just assessed across a variety of different hospitals and scanners, but is also assessed within protected subgroups, patients of different demographics. So, in this particular case, we're using very commonly used data, for example, the TCGA and other public datasets that have been used extensively in computational pathology literature to train these models. And then we're adapting those to data that came from MGH and the Brigham, but we stratified that data based on race, whether the patients had insurance or not, the income status that was extracted from the postcode, and a number of different other variables. This is a study from earlier this year.

And we basically want to assess that any of these shifts that are commonly known, whether it's acquisition shift or just demographic shift, leads to any bias and disparities. And we know that acquisition shift, for example-- Here we're showing three different examples of the same slide being scanned from a different scanner, or in the case of radiology slide, the same patient being scanned at two different hospitals. And these cohorts that are quite often used for training do have large disparities in where the data was collected and what the demographic makeup looks like.

So, our approach was that we would basically look at every component of the computational pathology pipeline where we train these models in a weekly-supervised

manner and assess for every modeling choice and the impact that it would have on these disparities. This includes from different pre-processing steps to the kind of features that are extracted to the aggregation profile and using commonly used techniques for bias mitigation, for example, using contrastive approaches or other modeling choices that are known for bias mitigation. And overall, we found that the feature extractor is the most important component in improving fairness for these models, and this was assessed for basically every common modeling choice in computational pathology.

So, I'll stop here in the interest of time and I want to leave you with this interesting poem that Judith Prewitt, who is a real pioneer in analyzing microscopy images using computational tools, wrote. She was doing some very exciting work in the 1960s and in 1970s at the NIH and then at the University of Pennsylvania. She writes something that I think is very relevant today. So, "Optical illusions can deceive the subjective eye, but objective measurements and algorithms are assumed not to lie. It's often said that medicine could use such objectivity and thought that this justifies machine intelligence activity. Artificial intelligence is another craze that uses computers to cope with the diagnostic maze. Though the criteria for intelligence has never been resolved paper after paper claims that the problem has already been solved." So, I think it's very important to keep in mind that there's still several challenges that we need to go through, and they're mostly related to evaluation and assessment of these very large models. So, I'll stop here and thank you so much.

Stakeholder Perspective – The Considerations for Multimodal Foundational Models & Generative AI Frameworks in Healthcare

Dr. Bhatt: Dr. Mahmood, thank you. Thank you for all the work that your lab does in this area and thank you for being so humble about the amount of work you're doing,

but the importance of getting it right for clinical care. So, next we have Parminder Bhatia, who is the Chief AI Officer at GE Healthcare.

Dr. Bhatia: Good morning, everyone. This is Parminder Bhatia. I'm Chief AI Officer at GE Healthcare. I'm super excited here to talk about generative AI and its potential to transform healthcare. But just as important, we'll also talk about some of the technical guidelines and methods to manage the risk associated with some of this technology.

Firstly, before we get started, I would like to thank FDA for bringing this Committee together to discuss this highly relevant topic that has the potential to improve the quality of care for the patients, as well as reduce the cognitive overload for the providers. Just like FDA led way path for medical devices powered by deep learning to be safe and effective, they'll have a critical component in ensuring GenAI is used in a safe and effective way as well.

So today at the outline we'll be talking about three major components. We'll start by diving deep into generative AI and foundation models, what they are, why they matter, how are they different from traditional AI models. Then we'll talk about their healthcare impact. We'll talk about how foundation models empower providers to create tailored adaptive solutions through smart data analysis. I'm convinced that cloud and AI together have the power to reshape healthcare more in the next decade than we have seen in the past century. But let's not forget we'll have to-- We are dealing with people's health. So, we need to implement these technologies that are safe and making sure we are thoughtful and respectful as well as responsible. So, I'll spend most of the time today talking about frameworks for managing risk associated with generative AI as well as these multimodal foundation models.

Let's start at the very beginning. A couple of years back with ChatGPT, which was one of the first technologies to showcase the incredible potential of generative AI.

To understand ChatGPT, we break it down into four major components. The first is the transformer architecture. We won't go into the details of the architecture, but this architecture lets capabilities or components like ChatGPT to understand and generate human-like language by recognizing complex patterns, relationships, and context as we saw in the previous session as well. Secondly, the advances in compute bar have allowed us to process huge amounts of data in parallel. Thirdly, alignment instruction or human reinforcement learning shapes generative AI by using human feedback to refine models' output for accuracy, relevance, as well as looking into the alignment with human values. As we'll see today, having humans in the loop is essential for rolling these technologies out responsibly. Finally, we'll talk about foundation models, which are the backbone of generative AI, and that's where we'll focus next on.

Foundation models at the core are trained on vast amounts of data, and that is often unlabeled, giving them a strong foundation to tackle a wide array of tasks without requiring significant datasets upfront. We are seeing these models make a huge impact where models like ChatGPT, Claude, for language processing as well as are also used in healthcare for a wide variety of medical imaging and multimodal tasks as well. These models started as language, but one of the key powers of these technologies is they were multimodal. They always started with multimodal. So, that's what we have seen over the last 18 months. It started with language. Now they're adding voice, they're adding images. We looked into the previous example how easy it was to even add pathology into some of these components as well. And because these technologies are built on diverse datasets, they are incredibly adaptive. They can be fine-tuned for specific tasks, whether it's answering complex questions, generating unique content, or detecting conditions in medical images.

To give an example, at GE Healthcare as well we have been doing research into foundation models in X-rays and MRIs, which could be game changers. For instance, these foundation models could help analyze images faster, more accurately, generate automated reports that support radiologists in their daily workload. Another major advantage of these technologies is that they can handle different kinds of datasets. It could be voice, it could be images, it could be text. Both at the input level as well as at the output level. This versatility to handle multimodal dataset at the input level as well as at the output level is kind of game changing and as well as invaluable, especially in complex fields like healthcare.

Next, talking about foundation models and how are they different than traditional AI models. Foundation models represent a quantum leap over traditional AI models in three key areas. Firstly, traditional AI models rely on limited and narrow datasets. Foundation models, however, are trained on vast datasets, capturing a wide range of knowledge. This results in actually more reliable and insightful predictions. Older models are often defined or designed for a single task. It could be text-only or image-only processing. However, foundation models being multimodal can actually handle multiple types of datasets simultaneously. This capability opens up new possibilities for richer and more contextualized AI applications. Finally, traditional models are usually built for specific tasks and lack flexibility to adapt to new contexts. This can lead to longer development cycle as well as massively increased cost for the healthcare cont. Foundation models by contrast, bring broad adaptability, where they can be fine-tuned for various applications, making them more cost effective and impactful across multiple domains.

I'll give an example over here of how foundation models can transform breast cancer care pathways. Today, patients move through a complex workflow. From

screening to diagnosis to treatment, as well as to the monitoring, where they're going through various devices and components such as ultrasounds, biopsies, MRIs, and PET scans. Because of the challenges associated with traditional AI that we just talked about, we see various challenges that can arise. For instance, that can lead to inconsistent screening adoption, complex biopsies, as well as various other challenges. It can also lead to low trial enrollment, which would slow innovation, while it can also lead to treatment selection, guideline adherence, and follow-ups, which remains inconsistent. Foundation models being flexible, multimodal, and adaptable can actually help streamline this fragmented process addressing these key pinpoints with data-driven, precise workflows that improve patient outcomes and care efficiently.

Now, let's bring all of this together and explore how these innovations can actually translate to tangible outcomes that can actually improve throughout the patient journey. Traditional machine learning crafts data from each clinical modality in separate silos, which creates significant inefficiencies. Foundation models, however, break those barriers by synthesizing complex multimodal healthcare, longitudinal healthcare data, from diverse sources into a unified view. These models are highly adaptive. So, over here we can see an example and we can actually think of an oncology-focused foundation model for breast cancer, for instance, can actually be adapted to other cancer related areas like prostate cancer. They can even go across multiple care areas, where a model from oncology can go into cardiology or neurology. Take Alzheimer's, for example. Imagine if a new drug receives regulatory approval. Foundation models could help develop the new care pathway solution in the matter of weeks versus months or years that it takes today, revolutionizing how quickly we can adapt advancements in the field of medicine.

Next, we'll talk about some of the considerations and measurements that we need to ensure to make sure we can actually bring some of these technology-- GenAI-enabled devices into the system. To ensure the responsible rollout of these technologies, I'll focus on four major pillars. Firstly, we'll discuss about the defined intended use, which is critical. By clearly outlining the product, what the product is designed to do, we can actually set the foundation for every subsequent step, from design to regulatory compliance. This clarity ensures that product aligns with the user needs and industry standards. Secondly, adopting robust risk management practices is essential. By identifying and addressing potential issues early, we can enhance safety, improve reliability, and build trust with users and stakeholders. Premarket evaluation comes next. This involves rigorous testing and validation before the product goes to the market, catching in on potential problems and ensuring it meets the quality and regulatory standards. And finally, establishing change control and lifecycle management is key. These systems allow us to manage updates and improvements efficiently, keeping the product compliant, high performing, and relevant throughout its lifecycle. With these measures, we can navigate the transformative potential of generative AI responsibly and safely.

Next, we'll go mainly discussing two major topics, which is-- First is the intended use and then around risk management. The concept of intended use is fundamental to the FDA regulatory system. For a medical device, its level of risk depends on its defined intended use. This principle also applies to GenAI-enabled products. To scope the intended use of a foundation model or a generative AI, developers can take several approaches. For example, they can control the prompts to shape the system's behavior. They can create boundaries on what model can generate. This allows developers to focus the product on specific use cases. Defining use cases or

intended use case initially is going to be critical for ensuring the trust and effectiveness; starting with a specific intended use case and expanding it gradually. From the regulatory standpoint, the process starts with first FDA authorization that corresponds to the scope of the initial use case, complemented by predetermined change control plan, or PCCP, that allows us to use these use cases to expand over time with the same intended use case.

In this slide, we can see an example of how we might want to start with, for instance, cardiac segmentation in ultrasound using GenAI or foundation model tool, and later expand to segment structures in kidney or other organs. We can do this as the intended purpose and workflow remains unchanged. However, detecting and categorizing lesions might represent a new intended use case, as it involves interpreting clinical significance and influencing diagnostic decisions as well. A similar process can actually be used even for GenAI or agentic AI systems designed to fulfill a specific task as we seek to expand the universe of possible use cases.

Medical device manufacturers use established risk management methods by assessing the impact of failure modes in the context of use, as well as designing mitigations to address the possible hazards. The approach equally applies for GenAI-enabled medical devices as well. However, GenAI and foundation models also introduce new failure modes, like hallucinations, where systems can generate fault or misleading information. The good news is by anticipating these use cases, we can proactively design effective mitigations, ensuring safe and more reliable use of these advanced technologies. This structured approach helps systematically to identify, as well as to address, potential failures to enhance device safety and reliability.

Just like we would do with medical devices, we have to be cognizant of the risk involved in the rollout of these technologies and think about strategies to address them.

There are some of these components which are common for even traditional AI models. For instance, misrepresentation of information can happen even with traditional AI systems, and we have to be aware of these limitations. We can apply techniques such as explainability and transparency to address them. For GenAI systems, which can also produce hallucinations, ontology-based methods or reasoning, something these models are really good at, can ensure accurate and reliable information generation. We can address the risk. One of the common questions that comes around is the inconsistent output that these generative AI solutions can build out. We can actually de-risk this risk associated with inconsistent output with mechanisms such as temperature control, which can help us to ensure that the model output remains consistent and repeatable over time. Finally, across the board we find that issues can be mitigated by keeping human in the loop, allowing for essential oversight and judgment and critical decisions. The level of human involvement is dependent on the intended use case that we look across these different areas as well.

Now let's look at an example of how we might mitigate risk. AI models trained mainly on common diseases may struggle to recognize rare conditions, leading to potential misdiagnosis. For instance, a patient with rare autoimmune disease might be misdiagnosed with a common condition like rheumatoid arthritis due to gaps in AIs training data. A machine reasoning algorithm can actually help by enabling the AI to use reasoning, which can be both deductive and inductive approaches, which can be enabled by chain-of-thought reasoning to better navigate complex medical cases. This can be extended with ontology-based disease models that can provide a structured framework for organizing and sharing medical knowledge, which can improve the diagnosis, treatment, and research efficiency. So, by bringing together machine

reasoning and ontology-based models, we can actually enhance AI's capability to manage both the common as well as the rare conditions in healthcare.

Now I want to give another example and another technique that can be used to mitigate risk. Visual grounding, which involves locating relevant objects in an image using natural language descriptions, enhancing image interpretation and diagnosis in medical context. This technology improves image understanding by allowing for more precise identification of abnormalities. In the example of this slide, a radiologist generates a report describing a condition observed in an X-ray, such as pneumothorax or lung consolidation. The report is processed by medical report grounding system, which highlights specific areas in X-ray that correspond to the descriptions, assisting the clinician by visually marking relevant regions and ensuring consistency in interpretation. One could imagine the same visual grounding system to also apply to the output of a foundation X-ray model generating a report that links directly to the content in the image, kind of creating a flywheel around that as well, which kind of eases the review by the radiologist. This is just one of the many examples and there are many more which we can't cover with a shortage of time, but we'll be covering more of these mitigation techniques in the docket that will be available in coming weeks as well.

One way to leverage the power of AI is to select the foundation model that best aligns with your development needs, and then fine-tune it using domain-specific data for a specific task. For example, with SonoSAM, we started with Meta's promptable image segmentation model, initially-- Which was initially trained on 1 billion segmentation mass from natural images. We then further tuned this model on 200 K more ultrasound image mass. In this case, we didn't know if SAM was trained on ultrasound images or not. So, our first approach was to understand its baseline performance. That's when it was determined that fine-tuning would be essential in this

use case. We eventually compared the performance of SAM and SonoSAM on an independent and representative dataset of ultrasound scans to understand the behavior of model on a variety of clinical images. It's important to note that the level of information about the foundation model or dataset used for pre-training does not significantly impact the process used for validation. We validated the fine-tuned model on an independent, representative and diverse dataset against the ground truth to ensure adequate performance as outlined in this paper as well.

As I near the end of my talk, I want to take you through three-step framework to think about change control and device management for GenAI-enabled medical devices. Firstly, change control, which is maintaining appropriate version control and end-to-end validation, ensures that GenAI-enabled devices produce consistent and reliable output. Adaptability and speed. We just saw that by leveraging PCCP, or pre-configured controlled plan, we can safely expand the model's use case within its intended scope. This setup can also be used for quick iterations, such as adopting new foundation model versions, to keep pace with advancements without compromising safety. And lastly with post model monitoring by applying best software practices are essential for capturing feedback during real-world use, and we have seen some of those today and we'll share more of those in future as well.

So, in conclusion, I hope you got the sense of how excited we are about foundation models, generative AI, and how they can have significant potential to transform healthcare. With the right controls, this can effectively be integrated into medical devices to enhance their functionality. Leveraging existing best technological practices within the current regulatory framework allows us to manage these integrations responsibly. What's more, combining foundation models with PCCP enables us to implement changes safely and efficiently, ensuring timely advancements

while maintaining high safety standards. Thank you for your attention today. Thanks for your time.

Stakeholder Perspective – Measuring Performance of Generative AI – Methods and Lessons Learned

Dr. Bhatt: Thank you very much, Dr. Bhatia. Our last presenter for the morning is Dr. Pranav Rajpurkar, who is the Assistant Professor at Harvard University. Dr. Rajpurkar, please, your presentation. Thank you.

Dr. Rajpurkar: Delighted to be here today. Thank you so much for having me. My goal today is to share some of the lessons that we have learned in trying to evaluate generative AI applications. Here's my disclosure.

There are many applications where we can have generative AI make big impact. One of them is the ability to generate text from images, be able to summarize clinical notes, be able to record conversations between patients and doctors, and to be able to summarize them. It's not just about text outputs. We can enhance images, we can generate visualizations, which can all be helpful to different stakeholders. A few months ago, I was shadowing a GI colleague of mine at one of the Boston hospitals and they use these narrow AIs to be able to highlight colon polyps on the screen. There was a time at which one of these detections came and the GI doc didn't know what it was about. Wouldn't it be cool if we would be able to ask an AI model a question, explain the object appearing on the screen, and the AI model could understand the intent and be able to answer that query in real time? These are the kind of technologies that I'll be talking about today.

The reason that we are seeing a lot of advances in generative AI is because of technology advancement on three fronts. One of them is the ability to leverage on-label data; another, the ability to leverage different kinds of data or multimodal data; and then

the ability to communicate that in natural language. So, the main question I want to talk about today is if we have a generalist AI system that can generate these text outputs for a variety of inputs, then how do we think about evaluation? And I want to break it down into two fronts.

One is how do we think about automatic metrics which allow us to do things at scale. And the second is what is the role of humans in the loop, of experts in the loop, as part of this evaluation. So, I want to start with the question of metrics. What metrics allow us to determine, in this case, whether a report that's written by an AI matches a report that's been written by a human?

Here are the examples that we can see on the left and we're going to look at different metrics on the right. So, we started working on this problem in 2020, and one of the literature fields which has looked at this question a lot, has been the field of machine translation, where we have translation produced by a model and then a translation written by a human. And there one of the most common metrics are metrics like BLEU. And what BLEU does is it looks at word overlap to determine similarity. For example, here there's a perfect match on certain terms like air bronchograms, and this is reliable when identical terms are used. However, this often fails where different words are used to convey the same meaning, so "left lower" and "left base," "consolidation"- "opacity," "effusion" - "fluid collection". This is not something these can recognize. So, how do we solve this? So, the time we turn to thinking about embedding-based metrics. Embedding-based metrics decide how to embed every word such that we capture the meaning of that word. Here, "consolidation" and "opacity" would be considered similar; and we were using this metric called BERT Score. "Effusion" - "Fluid collection" now becomes similar. However, one pitfall of these metrics that often happens is that opposite meanings can be represented very similarly.

So "with" and "without" mean two very different things. We want them to be far apart, yet these metrics often don't distinguish between the negation and the presence of something.

So, in 2020 we started to think how do we design a metric that captures what matters in clinical medicine. And so, we came up with this clinical accuracy metric called CheXbert to evaluate the medical context within the text rather than the text itself. So here understanding things like "with and without mean opposites" correctly, mapping synonyms to standard terms, preserving the negation, are all the positive things. However, this took us several months to build and was limited to a set of limited conditions. We also could not link these findings to the locations in which these findings were occurring. So, if you have a left pleural effusion versus a right one, this is not something this metric would be able to help us disentangle. So, how do we solve it?

So, we started to think how can we extract medical knowledge from these generations. And we developed a metric called RadGraph-F1, where the idea is: each one of these reports contains medical context; can we explicitly extract the medical knowledge here? And so, we generate these graphs off of these two reports like you see, which captures the different entities and the relationships of those entities to their absence or presence, but also to the anatomical location. And that's very helpful. Now we can compare these two graphs with respect to each other. The advantage of this is that we can capture a bunch of findings, we can preserve the relationship between the finding and the anatomy and we can handle negation, but this is quite hard to extend across modalities. RadGraph-F1 was developed for chest X-rays. We collected a bunch of human labels; we created a model that would be able to automate part of that. And then the second challenge is it's very hard to normalize across entities to be able to

understand that one medical term is actually a subset of another medical term. It's a challenge that's very hard to solve.

So fast forward three years later, we have been working very actively on this front. A metric that we developed, called HeadCT-One, is able to compare two reports using knowledge ontologies. Here, the original report and an AI-generated report have both been tagged with the anatomies, the descriptors, and whether or not an observation is present or absent. So, what you can see on the right is we can now understand that the extracted entity of "brain" and "cerebral" mean the same thing. "Atrophic changes" and "atrophy" mean the same thing. And the "small vessel disease changes" means the same thing as "microvascular ischemic changes". And we understand that the content of these two reports, where it matters, is the same.

But I think one question that we should start asking when we think about automated metrics is, "Does this align with what we care about?" In particular, if we asked experts to score this report to say, "This is the AI report, this is the generation. Do you think this is correct?" And they say there's one significant error here, there's one insignificant error here. Do any of these metrics allow us to really get at that salient question? So, how do we design this experiment? We thought in 2022, "Let's go ahead and understand the correlation between each one of these metrics and the expert rating." And so, we have the radiologist score on the x-axis and the metric score on the y-axis here, and we're going to evaluate these four metrics. And here's what we found. We found that the expert scoring revealed the highest alignment with RadGraph-F1 compared to BERT Score, CheXbert, and BLEU. But what we also found is that these different metrics are really capturing different ideas, and maybe it makes sense to think about how can we get the best of all worlds in creating a really good metric.

And so, we thought, could we compose these metrics? Could we take the weighted combination to be able to create the alignment with experts? And that's what we did with the metric that we proposed called RadCliQ, where we showed it has the highest correlation to experts and actually combines these different metrics together.

But often—and certainly is going to be the case clinically—a single number that tells you how well a report matches another one is often not fine-grained enough to tell us what exactly is wrong here. And we believe there is a strong need to understand the source and the type of the errors in the generations are made. And we break it down into two questions. One is "What parts of the generations actually have errors?" And two, "Is this a significant error or is this a stylistic one that won't make a difference to patient management?" And so, we developed this metric called FineRadScore.

But before I talk about the metric, I want to talk about the idea here. The idea is "Let's start with an AI-generated report and let's ask experts to fix it line by line." Now, they might decide to remove certain lines, they might decide to edit certain lines, and they might decide to add certain lines. Each one of these edits has a cost associated with it or a level of significance associated with it. For example, the first two errors might be non-actionable or actionable non-urgent errors, while we might have an emergent error. And we really care about catching those emergent errors; we don't want to make worst-case mistakes.

Here's a comparison of what happens when you look at the average error made by an AI generation and a human. If you look at the different error severities on average, you can see that if we take a look at something that's a clinically significant but non-emergent error, which is at two, 68% of AI reports and 69% of human reports have errors of two or less. And this seems great; it's almost equivalent. But it doesn't tell you, and doesn't tell us, the full story here, which is if you don't look at the average

error, you look at the maximum error, what is the worst error that is produced by the AI? You see that there's actually a significant difference between the AI generation and the human generation. And this goes to show and say that we should not be thinking about averages. We should also be thinking about worst-case analysis.

One of the things that our lab is actively working on is "Can we automate this process of fixing AI-generated reports given access to the expert reports?" such that it's not something we have to go out and seek experts for in order to be able to do and FineRadScore performs this task pretty well.

So, I want to switch gears a little bit and talk about current state of the arts in terms of building these generalist models. And I want to share one work from our group called MedVersa. The idea of MedVersa is to think about building a model that goes across different modalities. Typically, we have different models that work on chest X-rays, on dermoscopy images, on CT images. Would it be that crazy to have a single model that could look at all these kinds of images and be able to perform a host of different tasks? These might be tasks like the ones we've talked about today, where we're generating a report, but it might also be traditional tasks like being able to classify a disease off of the image or even being able to outline the area of interest, draw a bounding box.

What we found was that when we trained a single model to be able to do these tasks, we found that MedVersa was able to outperform the previous state of the art on medical report generation—we're here using RadCliQ;— on being able to classify images—we're using the F1 score;—on being able to draw bounding boxes—we're using the Intersection over Union. And as we've seen earlier today, lots of images are used to be able to actually train such a model. And this idea that we are working towards generalist models that can actually perform better than the specialist models is

fairly new, because it's been long assumed that if you want something that works really, really well for this specific task, we should only go ahead and train for that specific task. That no longer seems to be the case.

So, I want to hopefully show a demo here of this model. So, I want to show here a demo of MedVersa as it goes through and generates a report on a chest X-ray image. You can see it outputs the output here for this image that should hopefully be visible now. One of the things we expect these models to do is look at the context. So, here it's a 70 to 80-year-old female, and we're going to generate a report here. One of the findings is moderate pulmonary edema. We should be able to ask a question particularly on that finding, so to say "How severe is the edema?" And it should say an answer that's consistent with the report that it's generated. Now we might want to do a task like being able to outline the heart. It's not quite right here, but I like to say its heart's in the right place.

We'll have to think about how to evaluate such systems, but it can look at dermoscopy images, be able to make a diagnosis, be able to perform tasks like segmentation on 3D modalities—here being used to segment out the liver,— and having a new task be solved, be as simple as simply changing the prompt used in this particular case.

So, that's MedVersa. And I wanted to share this example as an example of the capabilities that we're thinking about when we're talking about the next generation of multimodal generative AI models and what it takes for us to evaluate them. So, I want to now talk about human-centered evaluation, because that's such a key area. And one thing I'll mention before we do that is if we look at automated metrics, what I just showed you today, MedVersa is currently the best benchmark model in terms of generating reports on chest X-rays. If over the course of the next year or two, I expect

we'll see significant progress here and benchmarks are going to be very important. And we have the ReXrank benchmark that we have generated for this now.

So, onto human evaluation, how do you leverage expert capabilities to be able to evaluate such a generalist system? Well, one of the ways that we thought about it was to say, "We're going to ask experts to choose which report is better." One of them is going to be generated by another expert, one of them by the AI model. Now, the expert might say, "Both of them are equivalent clinically," or might say, "Report A is better" or "Report B is better." What we found was that overall expert-written reports are preferred; AI-written reports are preferred some of the time; and they're equivalent more than 50% of the time. We have a breakdown of how things perform on abnormal cases as well as normal cases. And on abnormal you see the human preference being greater, and for normal you see they're pretty much equivalent most of the time, and an equal split of the remainder between whether the AI was preferred or the human was preferred.

One of the other ways in which we think about evaluation is to think about what is the setup in which they're going to be used in practice. One of the exciting opportunities here for radiologists is to be able to spend less time on cases. The idea of working off of a draft and being able to edit it to sign off is not new. That's how academic radiology often works. And so, could we have a system in which these AI drafts are given to experts for them to modify and then sign off?

The way we set this up is we say, "We're going to measure the time spent on a study, the mental effort and the confidence in each report as radiologists are going through X-rays." The control here is going to be starting off with a normal negative template and the AI model as well as the radiologist is going to be supplied with the demographic, the indication, and also the comparison.

And what we found here in a small pilot study right now is that there is, in fact, significant time savings and mental effort reduction, as well as increased confidence in each study. But the caveat here is that it's going to be very important to think about how this technology can work for everyone. Even in small-scale trials, we're seeing some people experience big benefits from this technology while others show no improvement in their performance or their time taken as a result of these technologies.

I want to switch gears a little bit and talk about one of the things that we've all been very excited about. It's the ability of large language models to be able to have conversations. Now, there's a big disconnect in the way that these systems are evaluated and the ways in which these systems are expected to be used in practice. And I want to share one example of that. The primary evaluation for LLMs right now is based on answering medical exam questions, and we've seen incredible numbers in terms of the ability of LLMs to outperform humans on this task. Several studies, including the ones this week, that even show that the AI alone performed better than the clinician being assisted by the AI.

But these exam questions and the setup is broken in fundamental ways. One is that no patient ever presents with "Here's a summary of my symptoms, here's my vignette;" no patient talks about their medical terminology in such precise terms; and no patient gives options for a doctor to choose between. And these are ways in which we need to change the way we're thinking about the evaluation of these clinical large language models. One of the ways in which we're thinking about this is "How can we have benchmarks for conversational assessment of LLMs?"

Now, we can get these LLMs to have conversations with lots of patients, but that would be unethical in many ways. "Could we have LLMs simulate patients to have these conversations with clinical LLMs and see in that setting whether or not these

clinical LLMs do well?" is how we're thinking about this. So, I want to present our results here on evaluating clinical LLMs on 12 medical specialties, both text-only and multimodal LLMs, and talk a little bit about continuous monitoring care.

The setups we have is, we have these case vignette evaluations, which we know are unrealistic, and what we do is we take these case vignette evaluations and ask a patient to share information with the clinical LLM. The clinical LLM asks questions to the simulated patient. The patient answers using the available information in the vignette and we evaluate this in a multi-turn, in a single-turn, in a summarized conversation setting. The main point was this, that when we actually get these clinical LLMs to have conversations, to ask the right questions, they actually do much, much more poorly than when they're given the vignette. A large part of this has to do with incomplete gathering of relevant medical history for these clinical LLMs to actually have all the information to arrive at the right diagnosis. Similarly, multimodal LLMs are severely limited in their image interpretation capabilities and this becomes worse when we try to evaluate them in this conversational setup.

I want to end with talking about an application of generative AI that has several different components, but goes to show the potential value in terms of helping patients understand their own medical information. A lot of us, or our family members might have had experiences in which we get scans and we get radiology reports associated with those scans, and the scans come on CDs and the reports come over email, and it's very hard to understand what any line of the report means, especially for those who might not be medically trained. So, we thought about, "Could we solve this problem by creating a system which allows us to understand different lines in the salient information in the radiology report and in particular be able to link it to regions of the image and also regions of the anatomy?" This would allow a patient or their family

member to understand what's going on in their image, what was found, what does it look like, be able to compare to what a normal scan looks like, be able to see the organ and also have this virtual avatar do some explaining. Now, this is a system that can be built with a combination of existing components and components which we developed to segment out abnormalities. I now want to play you a demo of what this system looks like.

Okay. So, in this particular setting, there's no audio, so I'm going to be the virtual avatar in this case. But, the first finding that's being explained is that of aortic calcifications and as the system scrolls through the aorta, you can go ahead and have the virtual avatar talk about this. We are not going to go through the whole video, but if we fast forward it, we'd also have a comparison to the reference normal image, right there, and then also the organ rendering on the right. And this is our vision for what might be a world in which we have the patient be empowered to make their own medical decisions.

We can go back to the slides. So, in conclusion today I wanted to share some of our lessons of evaluating LLMs and generative AI. And what I hope to have shared with you is that they're challenging, that there are automated metrics that take us a long way, that there are human-centered evaluations that take us a long way as well. The space of opportunities is immense and it's a tremendously potentially helpful technology, and we need to think about the best way and the most safe way of bringing it to practice. I'd like to thank my lab members who made all this possible. Thank you.

Dr. Bhatt's Summary

Dr. Bhatt: Thank you, Dr. Rajpurkar. Surprisingly, and I want this formally noted for the record, we are ahead of time. Thank you very much to our speakers in the room. I'll take a minute to just go through the whirlwind we just heard one time in summary and then we'll go to a break. I think it's quite a feat we have ahead of us, but we can

already see that we have really talented people we can rely on, and so it feels good to have had this morning happen.

We started with Dr. Califf, who reminded us it's imperative to bring technology to those who need healthcare the most and have limited access. But then, he specifically said the criteria for adopting AI are largely financial and we need to think about clinical outcome measurement. I think that our lectures today helped us in that. He also challenged us that we don't have an information ecosystem. Local ongoing validation of AI is something we need to focus on.

We then had Dr. Michelle Tarver and Troy Tazbaz speak with us. And they reminded us that we can have a collective conversation to move forward equitably, that we need to think about AI for access. It's interesting. They've mentioned home health quite a bit from the CDRH, and when I was training, we used to talk about the last mile of healthcare, and that was the patient in their home. And I think what I'm hearing is we really need to focus on the first mile of healthcare and how we're going to diagnose patients earlier and help them earlier. So that's part of the conversation we're here for today.

We then had Dr. Badano speak to us, and he talked about regulatory science challenges. He talked about the challenge of defining scope, the fact that foundation models actually don't live necessarily with a device manufacturer, how oversight is important, hallucinations must be controlled and diverse data needs to exist and are sometimes challenging to find. And then reminded us that performance assessment has been established. We know what benchmarking looks like, we know what expert evaluation is, and now we are adding model-based evaluation, where there may be some challenges like inter-model leakage, but also a lot of potential benefit.

Dr. Mahmood next gave us a real impressive look into pathology and how far ahead they are. He reminded us that input patching, feature extraction, feature aggregation and prediction is a model we've always used, but that feature extraction is actually where AI can make a significant difference, and showed us what self-supervised models look like. Very impressive to see OncoTree go from 43 to 108, but really humble in saying, not necessarily ready for clinical use when we expand that much, that we still have work to do. So, thank you for that.

Interesting to see clinical reports versus generated reports in both his presentation and the presentations following and start thinking about what that looks like when we not only talk about the models and how they're going to work, but how we're going to understand their output. Because it's not only whether that model works, is whether or not we can do something clinically with it that is also correct. We then had the opportunity to hear about PathChat. Thank you for thinking about equity and global equity, not just United States equity.

Moving forward. Important to think about why foundation models. And Parminder Bhatia talked to us about training. How we can go from narrow to large training sets, modality, the ability to do multimodal analysis that didn't exist before other than in the human brain, and adaptability. Important to think about intended use, robust risk management. And thank you for bringing up the FDA's pre-control change plan. It is an excellent way that people have been able to look at devices. It may be applicable as we think about generative AI in those devices.

Dr. Rajpurkar gave us an excellent look at what we are doing now for performance measurement in generative AI. I think it's so important to recognize that we can leverage unlabeled data, that we can leverage multimodal data as we saw. And again, coming back to communication, how do you communicate that in a natural

language way that those who are going to use these tools will use them as correctly as the tools are able to work? I think as I look at all of this, one thing I will ask people to do is to think, importantly, but not talk about during the 10-minute break, "How does this help us think about premarket evaluation, ongoing risk management, and postmarket surveillance?" Because we've heard some very different fields, but we recognize that each of them applies to what are we doing premarket to assure a right? What are we doing in terms of risk management? And what are we doing when we are thinking about postmarket surveillance? So, just while it's fresh in your head, sometimes thinking about it by yourself.

We will take a 10-minute break and maybe we can take more time for lunch later. We'll take a 10-minute break now. Panel members, again, joking aside, please don't discuss the meeting topic during the break amongst yourselves or with any of our panelists and guests yet. We will have the chance to talk about this when we come back from the 10-minute break. Lastly, yes. I'm going to be writing with pen and paper all day. I realize this is a Digital Health Advisory Committee, but this is how my brain works. So, as you see me scribbling, some things maybe you're never going to change. Thank you and enjoy your break.

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

Dr. Bhatt: Well, I would like to thank all of the speakers for presenting this morning. We now have an opportunity for Open Committee Discussion and clarifying questions from the Committee. As a reminder, although this portion is open to public observers, public attendees may not participate, except at the specific request of the Committee's Chair. That's me.

Additionally, we request that all persons who are asked to speak, please identify yourself each time. This helps with transcription. So, for the panelists, please just

quickly say your name before you have your clarifying question, and then when you ask it to one of our five speakers this morning, if the speaker before you answers that clarifying question, could again say your name and then answer the question? That will help our transcription process.

With that-- Again, a reminder, if you have a question, you can either raise your hand or sometimes it's just easier to put your card up vertically like this. And James and I will keep track of the order of questions. And we have about 30 minutes for this process. Panelists, all yours. Okay. Dr. Apurv Soni.

Dr. Soni: Hi. Apurv Soni from UMass. One of the points of consideration that I've been thinking about as we looked at different approaches for evaluation of generative AI is, "Where does AI assurance labs fit into this ecosystem?" and "What role do we imagine AI assurance labs having?" And I'll just kind of add my perspective on it. I think there has been a lot of momentum behind Coalition for Health AI as an organization that has grown. And there is also a lot of emerging momentum behind development of AI assurance labs that may allow for either human-in-the-loop evaluation or model-based evaluation. And how does that interplay with some of the metrics that need to be evaluated for FDA's determination.

Dr. Bhatt: You didn't want to address that to anybody specific. So, I will now maybe say, do any of the Panel members, or more importantly, do any of the five speakers who perhaps have been thinking about human-in-the-loop, have a thought about that? Otherwise, we can take this as a comment as well for later. Okay. Such an important area to think about with AI assurance labs, and where they work. Let us continue with questions for now, and I think we can bring that conversation in a little bit later. Thank you so much, Dr. Soni. Anybody else? Yes, Dr. Shah.

Dr. Shah: Hi. This is Pratik Shah, University of California. So, my question is specifically directed at Dr. Aldo and Dr. Mahmood, or Dr. Rajpurkar, either of the three.

So one is that when we have GenAI models, we know that they will generate information that they necessarily were not trained on for clinical decision-making. And at the FDA, we have three different pathways. We have the De Novo pathway, premarket approval pathway, and then we have the 510(K) pathways. Right? And for the top two, we have those pathways activated when there is no other predicated device in the market for those pathways. So, my question is, when you have these kinds of models where there are no predicated devices in the market and they're making decisions or determinations based on training data that they haven't necessarily been completely trained on, for example, if you have a weekly-supervised learning or multiple-instrument learning or statistical learning, it hasn't seen all the data and it makes a determination. What actual regulatory science pathways should be considered these models fit into? That's one sub-part.

And the second part is, "What is the best way to do a randomized controlled trial on some of these models?" If that's one of the ways we want to go. Maybe a human-in-the-loop model where the AI model doesn't do clinical decision-making and the human continues doing clinical decision-making. Or a second pathway where it's an assisted care model. Right? So, these are some of the bigger concentrations to Dr. Aldo or any of the speakers, Dr. Mahmood or Dr. Rajpurkar, I can take. Thank you.

Dr. Bhatt: Maybe let's take that second question first, right? When we talk about level of testing necessary for us to trust something. Randomized controlled trials come up in clinical medicine a lot, but how are we proposing to test things? Do any of the speakers have kind of a thought or two? Yes, please. Thank you.

Dr. Rajpurkar: I think of this in two ways. One of them is, "If we have a generalist device, could we scope it to tasks that we can evaluate really well?" As an example of this, you can have a multi-disease detection model that can detect a bunch of different diseases, and we can apply existing frameworks to be able to do evaluation of those narrow paths. So yes, there are a lot of them, but we can evaluate each one of those individually.

The second aspect is if we leave it in its unconstrained format where there's a bunch of inputs, there's a bunch of possible outputs, one of the ways in which we need to transition our thinking is from thinking about sensitivity, specificity for individual models to thinking about preferences. And this change to preferences has already been something that outside of medicine, people have thought deeply about.

And the idea behind preference-testing is the following. We have, in a lot of cases, reports generated in the clinical settings by people, and we have reports generated by the AI model. Let's ask a bunch of experts, "Which one of these do you prefer?" and "If you are preferring one over the other, is it because one of them would lead to clinically significant differences in terms of patient management?" Thinking about both the preference and the reason for the preference. We can set a non-inferiority margin to say that if we fulfill this non-inferiority margin, then this is at least as safe. And then downstream, we can think about it as at least as effective where the increased accuracy might come into play or the efficiency might come into play.

But I think this transition to thinking about preferences and evaluation on a dataset, which is representative of the kinds of settings in which we want to deploy, as well as with the kind of people that we want to deploy, too. For instance, if something is targeted at radiologists, then those should be the users. If that is particularly for residents, then those should be the users of that.

And I think we don't necessarily need to be in a setup in which we need a prospective double-blinded randomized control trial. In a lot of the settings, the blindness of the randomization won't even work because it'll be easy to see what's actually a normal negative template versus what's an AI generated report.

And I think we can actually go a long way, even with the retrospective evaluations to tell us, based on preferences, how well certain models perform. That's my two cents.

Dr. Shah: Right. Quick follow up, just to clarify. At the FDA, we have this intended use for medical devices where we want to label the product by the people for what the product was intended to do. For some of these things, which are so broad-- I think you answered, Dr. Rajpurkar, some of those questions. For some of these products, which are so broad, it may be hard to come up with an intended use, a priority. Because once you have an intended use, later on you can have tracking and then software product lifecycle where you can track the model's performance. So, I think one of the things that you mentioned is to actually have a very narrow definition of who the end-users are, what the product label should be, and how the performance should be evaluated in the software product lifecycle. Correct?

Dr. Rajpurkar: Yeah, and just to talk about that intended use piece-- I think if you're looking at, let's say, a modality and you're looking at the natural distribution of whatever scans are to be read, then that would be the scope of the usage. So yes, it's broad, but if in the intent that we're seeing retrospectively or with human usage, it's representative of the distribution that we want to deploy in, then that to me makes sense as a way to evaluate.

Dr. Bhatt: Excellent. Thank you. For the transcriptionist, that was Dr. Rajpurkar who was giving the answer as the speaker and the additional comment was coming from

Dr. Shah. What I'm hearing, just to clarify, is that when we're thinking about device scope, we are also thinking about who is using the device and the setting in which it's used and defining all of those are equally important and sometimes that could be limited in a randomized blinded mechanism. It does bring us forward to the question for later, which is "What does postmarket surveillance look like then if we are putting these into the market and then requiring an understanding of what is happening once they're being used in the right setting with the right people, and how can that be expanded?" Let's go to Dr. Peter Elkin next, please.

Dr. Elkin: Thank you so much, Dr. Bhatt. So, first of all, I'd like to say that for every-- To help answer Dr. Shah's query. For every model that would be even considered for approval, we need a model card that has on it the exact dataset in which it was trained, what the demographics of the people whose data was used are, and so that we have a sense of the breadth of the training. If there's an underlying foundational model, we'd need to have the properties of that model that's defined in the card about what is in the potential model. We need the results of evaluations that were done prior to this, including the sensitivity and specificity of the model for particular tasks. And we need to understand what those evaluations have been, how they've been done, what was the methodology that was done in the trials that led to the submission.

I personally believe, and this is my opinion, that in between a clinician in some specialty or a patient and a model should be a clinical informatician. Someone who understands how the model was trained with the inclusion/exclusion criteria, for the studies, what the sensitivity and specificity of the model for different tasks are. So that they can give good recommendations, interpretations of how they're used. Very similar to an echocardiogram; it's used by cardiologists, it's not used by every doctor, it's used

by specialists in the field. We do have a Board-certified specialty in clinical informatics. These people can be a link to help us to make sure that it's safe.

And then to answer the second part, because I have strong feelings about that too. There's two kinds of evaluations that we need to put in place. One kind of evaluation are for rare but significant errors that happen, so that we do biosurveillance of problems that come up with the model. And then, second is for drift. Over time, as medicine changes, as the population that's used on it changes, the accuracy and effectiveness of the model can drift. So, we need particular agents that are looking at this on a regular way.

So, we need to create machine learning algorithms, probably, not just human research capabilities, that can actually be constantly monitoring this kind of information and sending feedback to some kind of a central registry where we can have researchers, like the people who presented today so eloquently, look at the data and understand when we have a problem. And there should be Data Safety Monitoring Board for these, like any other large study that would, in a phase four trial, be able to bring it back to this group, or whatever the appropriate body is at the FDA, in order for them to gain some kind of headway of when they should do something to keep the public safe. In other words, have thresholds of action where, if the model drifted beyond a certain point, it would be recalled. Or if there were a specific subpopulation that was at pretty high risk of some secondary bad effect, that those would be notified and that the use of that model would be restricted not to involve people who are at high-risk for harm. And if anybody wants to respond to that, I'd be happy to hear. Thank you.

Dr. Bhatt: Thank you so much. That was very helpful. Dr. Kukafka, did you want to comment?

Dr. Kukafka: Yeah, thank you for all these wonderful presentations. And, it got me thinking, which I think was the purpose of these talks. And I just want to-- It might be-- We'll have more discussion in the afternoon about this, but as I'm thinking-- Because I do a lot of RCT type work in this space, and I just wanted to differentiate the way I'm thinking about some of the evaluation that we're discussing. A lot of what-- I think I've heard two things. A lot of what you just brought up or what I would term as process measures. And then I heard one reference to more endpoints, and the endpoint that I heard was around preference; the clinician preference, patient preference, I'm not sure, but preference.

I'm just curious at this point, and I know we'll have more discussion about this. What would be-- I think the process measures might be more well defined, I think. And I'm curious about what-- In evaluation, in a clinical setting, what would be the expected endpoints? Since it is very atypical to the very traditional kind of RCTs that we might have done in the past. So maybe just start to think about what would be the endpoints in this space? Not the process, but the actual clinical endpoints. And they don't have to be-- Preference would be one, because we do use patient preference, we use some decision-making metrics, which might be another clinical endpoint, like improvements in some clinical status. I'm just curious if there's any thinking about this area.

Dr. Bhatt: I love those comments. We have a deliberation section that will happen after lunch, and one of the ways that we were hoping to organize it, is to really think about structure, process and outcomes and how we're going to think about generative AI in that setting.

Just a quick reminder that I just got from our FDA colleagues, which is that we don't have this morning speakers beyond this morning necessarily. So, if we could use

our time to direct our questions to them, we will have more time for deliberation. And I've taken notes on all of those really good points.

If you could say your name before you start speaking, that would be great. And Diana Miller, you are up next. I did see that.

Dr. Radman: Thomas Radman. Thank you for asking your question on endpoints because I also wanted to ask some clarifications on this idea of preference-testing. And I'll give you my interpretation and then I'll invite the speaker to clarify. But I guess I wanted to say that-- So, we're here to discuss generative AI. And AI—let's call it maybe for the sake of this Panel, pre-generative AI—has been in existence for a while and there's some really great guidances and regulations that FDA has used to regulate that thus far. So, we can learn a little bit from the past to inform the future of generative AI.

So, when I think of preference-testing, and please correct me if I'm misinterpreting you and I have a question for you directly actually, but-- The FDA currently, essentially has two types of diagnostic devices. You have devices that directly provide a diagnosis, and I think you guys are allowed to correct me as well if I'm misspeaking, right? As the FDA. But, there's diagnostic and there's an adjunct to diagnosis, which is where I interpret this phrase of preference-testing because adjunctive diagnosis is-- The clinician basically is-- It's just helping him to-- And he could reject or accept the recommendation by the machine learning algorithm, pre-generative AI algorithm typically thus far. And then when we talk about preference-testing, you can imagine a trial where you have half of, let's say radiologists, if that's the example we want to use, use this device to adjunctively help them provide the diagnosis. Does it augment their ability to diagnose? And then you have half doing standard of care. And, so if I'm interpreting you right, what is preferred would be

whichever of those two randomized groups in a clinical trial performing the diagnosis better by some significant margin.

The question I had for you before you start speaking is, because we have this track record, if you will, of adjunctive diagnostic devices being cleared for marketing authorization by the FDA, what would be your recommendation for the number of users in such a study in, let's say it's a two-group study, with and without a generative AI device? And does generative AI, which should be our focus here, provide new questions above just a pre-generative AI machine learning device where you might want more users than the FDA is used to?

Dr. Rajpurkar: This is Pranav Rajpurkar. Let me try to be very concrete about the trial setup, just so we can have a working point. The trial, I'm imagining-- Let's start with the intended use case of drafting. So, there is a draft that's generated. A clinician will be responsible for editing that draft and signing off on it rather than writing something from scratch. Let's use that. So, the idea that we've had and performed at this point of a randomized controlled trial is one in which you have a certain set of clinicians that's starting from the current clinical standard, that might be from scratch. So, they have those and then we have another set that's randomly assigned to start with an AI draft and they edit and we get two final sets of reports that have been generated. Now, a panel that's independent of this reviews each of these reports and decides whether they preferred one that was generated with the AI human combination or we can do the classic human in their standard, or we can even have the AI by itself in the case that we're thinking about having autonomous generations. And then the preference is essentially what fraction of the time was one preferred over the other. And this takes into account, a lot of the time, clinically significant errors being made. If those are made, that shouldn't be preferred compared to the clinical standard. And so, the idea

behind a lot of this is to say we're not aiming for perfection, we're aiming for better than existing clinical standard and this is one way that we can think about setting this up.

The second part of your question had to do with how many users. And I think it's a very challenging question. Primarily because it matters what the diversity in the intended users looks like. And what we have shown in a natural medicine publication earlier this year is that things like years of experience, things like specialty, things like previous experience with AI tools were actually not good predictors of whether someone was going to experience benefit with the use of AI. So, it's unclear right now in literature what is the sample size needed from the number of users. One of the ways that we can set primary outcome is to think about the average. It's to say, "On average as a group, the clinicians that are representative of the intended users should experience improvement" and recognize that there is going to be a distribution where some clinicians are not going to benefit and some clinicians are going to benefit a lot by the use of that technology. How we make that variance small, we don't know, but I think that's going to be the reality of the evaluations today.

Dr. Bhatt: Thank you so much. We'll go to Diana Miller next and then Dr. Stanley after that.

Ms. Miller: Thank you very much for this session. I'd like to commend the FDA for organizing this and getting ahead of technology. As far as I recall, I think FDA approved over 950 AI-enabled devices as of past August. So, we did a good leap in that technology in the past and I'm sure we're going to be doing a good leap in GenAI technology in the future. So, this is a great start.

I have very three concrete questions to some of the presenters that are related more on how we design these models. So, first is Dr. Mahmood and computational

pathology. My question is, and maybe I missed that when you mentioned-- Sounded like you used a pre-trained model; what pre-trained model did you use at the beginning and how did you-- Did you have any criteria of selections of a specific pre-trained model? So that's question number one. How did you pick that?

And the second question is for Dr. Bhatia, and I'm unsure if I mispronounced the name. You mentioned that we use foundational models and we can transfer them from one domain to another domain, and we talk about the intended use a little bit and how do you define intended use in that space if you move from one domain to another. Maybe you can clarify that a little bit.

And then the last question is for Dr. Rajpurkar about metrics, and I know he's a popular one today, this morning. He gets all the questions, but one more from me. Did you look at our metrics outside radiology space and if not, how would you advise designing a metric for GenAI in other space?

Dr. Bhatt: Wonderful. Let's take those in order and start with the selection of the pre-trained model. Thanks, Faisal.

Dr. Mahmood: Hi, this is Faisal Mahmood. So, we developed our own pre-trained model. It was trained on a lot of internal data from Brigham and MGH, but also data that was sent from clinical collaborators at other hospitals. And we attempted to maximize for diversity just based on what we knew from conventional computer vision-based applications, that diversity of data is much more important than the quantity of data in training these large self-supervised models. And then we use that model for our generative AI tool. Thanks.

Dr. Bhatt: Thank you. Dr. Bhatia. Your question was to define intended use for foundation models if we think that there may be multiple potential uses.

Dr. Bhatia: Yeah. Hey everyone, this is Parminder Bhatia. Thanks for the question. So, I'll first actually try to just answer the first question that you just asked as well on how do you decide the foundation model. It depends on the use case that the developer or anyone who's working on it would look into. For instance, one of the examples I gave was ultrasound segmentation model SonoSAM that was built using SAM, which is built by Meta. And there you start evaluating it at the baseline, how is it performing on those benchmark datasets, which are more universal, and then you start to see "Is there a value to fine-tune that or use that model off the shelf?" If you go into multimodal models, you could even adapt it to models like ChatGPT or Anthropic and tune them for specific use cases as well. In some of the cases it might be beneficial, as I think Dr. Mahmood also talked about, training the model from scratch as well, the foundation model from scratch. So, I think it depends on the use case and the accuracy bar that you can get from these off-the-shelf models as well as trying to adapt from those use cases.

Now to your second question on how do you go from one domain to another. One of the examples which I would like to give is on summarization. You have data for cancer patients that could be thousands of clinical notes that can be created over the years. Now you could start with an intended use case on prostate cancer for summarizing that data. And now you can even extend as you're starting to think, "Can I add more domains into that?" Now, "Can I add from prostate cancer to breast cancer into those same use cases?", which is summarization. And same thing goes across on the report generation, as we looked into as well. So, you start with that. You're going from image, you're going into report generation, you start with a specific disease-based care area, and then you start adding more components and that's where you can actually leverage capabilities from PCCP and others to add those as part of additional components, which are again aligned with the same intended use case where you're

either trying to do summarization or report generation or even segmentation where you start with few organs, add more organs later as well. So, the intended use case is still the same. You're just increasing the horizon of how either you can adapt more organs or you can adapt for more disease care areas as well. Thanks.

Dr. Bhatt: Great. Thank you. Dr. Rajpurkar. And while you're coming up, so from Dr. Bhatia, just thinking a little bit about defining intended use, still very specifically in the area in which that was trained, the idea of the pre-control change plan maybe now applied to generative AI, which is a little different than the way it's used now, but might be something we want to pin for our discussion later. Dr. Rajpurkar.

Dr. Rajpurkar: This is Pranav Rajpurkar. The question was around the usage of metrics outside of the field of radiology. So, we've done an evaluation of generative AI models across 12 medical specialties, and one of the ways that we've set up metrics is diagnostic accuracy. And on diagnostic accuracy, in one of the setups, we actually have an LLM that's responsible for determining whether the diagnosis according to the model was the same as the one made by an expert. I think that's where the subtleties of, and the brokenness in some ways, of using medical exam questions as a way of evaluating these models comes in. Because our main finding was that, "Yes, you can have these LLMs, have these wonderful answers when they're given options, when they're given these very clear questions," but things break down once you think about them in a conversational setting. So, we have this benchmark called Kraft M.D., which essentially has patient agents that are simulated, that have conversations with these clinical LLMs. And the evaluation ultimately is across the cases; what was the diagnostic accuracy? And then we perform subgroup analysis to think about "Are there specific subgroups?", which are either age, sex, we could do ethnicity or even specialty where there are greater performances than others. We've also investigated this idea of

preference testing once again, where if you have a report that's generated, even on cases, by models and by humans, then how much do we prefer one versus the other?

Dr. Bhatt: Thank you. We'll go next to Dr. Stanley, then Dr. Rariy. I'll ask you to put your placards down once you ask your question to help me. Thank you.

Dr. Stanley: Hi, this is Laura Stanley. So, this is for the presenters. I'm loving this mashup of the academics with the industry folks and the regulators. This is fantastic. So, thank you. So, my question-- I might've missed this and I apologize if I did. So, I kind of sit at this human tech integration, human AI space, and what I wasn't hearing, and perhaps I may have missed it, I apologize if so, is this notion and there's a huge body of science around trust and automation.

Have you collected-- What are some of the metrics? What are some of the things? What are your thoughts about trust and automation? I spent 10 years working with autonomous vehicles. We have levels of automation which we might want to consider, if it hasn't already been considered, in terms of-- I think of this as like human-out-of-the-loop, human-in-the-loop, human-on-the-loop. That's a level of automation we might want to consider. And now, what about the trust? And when I think about trust, we have to think about "Are we over-relying?" I heard that come up. Over-reliance. "Are we vigilant? Are we situationally aware?" So, the need for human trust is critical in these machines as we're interacting. So, I'm curious from the presenters, if you collected that, what are some of your results? What are some of your metrics? There are validated instruments that do exist that people have modified from aviation, surface transportation, nuclear plants, social robots, etc.

Dr. Bhatt: Okay. Validated metrics for trust and automation when it comes to generative AI giving us answers of results. Yes. Dr. Bhatia. Thank you.

Dr. Bhatia: Hi, this is Parminder Bhatia. So, I think one of the areas where generative AI or foundation models is different than models we have seen in the past is explainability or what's called a chain-of-thought reasoning. So, these models are not just giving out a number which is positive or negative or confidence score, but they can also break down the problem and explain as to why they came out with the output as well. And I think that adds more to the explainability as well as trust as well.

So, you're getting the output which could be the summary and you can actually tweak it in a way to provide reasoning as to why it generated that output as well. So, I think that's where some of these technologies being multimodal as well as coming, thinking more from how a human would think, both from cognitive side of things and inductive side of reasoning as well, breaking down a bigger problem into smaller problems and coming out with rationales for that. I think that's one of the areas where generative AI can actually help decipher some of these things and make explainability more as a norm than we have seen in the past as well. It can actually be used for traditional AI systems as well. In some ways, we are starting to see research where generative AI or LLMs are being applied for evaluation or reasoning even if the output came from some other system as well. So, I think that's one of the areas where it'll help to bring in more trust into these systems as well. Because at the end you still have human-in-the-loop, but you're providing not just the output which is positive/negative, but reasoning behind that as well.

Dr. Bhatt: Wonderful. Thank you. We have five questions remaining, 10 minutes remaining. So, my panelists, if we could make our questions directed directly at the specialist who spoke. Dr. Rariy, then Dr. Maddox.

Dr. Rariy: Hi. Dr. Rariy. Thank you so much. As I trained within the Harvard system, in the Brigham system specifically, so it's great to see so many colleagues here.

One question I had was for Dr. Mahmood, specifically as a physician. One of the questions that I have, and I'm very curious to hear your response to, is around the end-user information or end-user labeling. So as an example with the PathChat, the information that's provided back to the doctor I think is as important in terms of the references as an example, or in clinical practice. I'm an endocrine oncologist, so I focus a lot on precision medicine and the information obtained back from that includes references, includes specificity, sensitivity, accuracy, and that very much impacts how we're presenting this information to patients.

And so, while as an example, ChatGPT, when we're asking questions, it just provides the answer back. There's not really that rigor or that layer underneath around "Where did this information come from?" or "What specific information from?" as an example of that pathology impacts the sensitivity or the specificity or how the user uploaded that information, whether it's magnification changes or things of that nature. So, I'm curious to hear from your perspective and recognize you with the understanding that it's still in research phase and not necessarily ready for end-user prime time. But again, with that physician perspective, that's going to be pretty important. Dr. Elkin already stated that the necessary labeling for FDA approval is required, but I would say and articulate that from the provider standpoint, that's going to be extremely important as well.

Dr. Mahmood: Yeah. This is Faisal Mahmood. So that's a very important question. So, there are a couple of components. The first one being that how do we quantify-- "How well the model is essentially doing?" The other one is, "How do we get to the underlying references where the information might be originating?" In terms of quantifying how well the model is doing or the model generating, how confident it is in a said prediction that can then lead to as a sort of a trust-score for the physician. It's

currently slightly difficult to do that directly from generative models. It's just based on where the machine learning research is. But what we think would happen down the line for these very specific generative models, for example, one that's focused on pathology, we would have a generative model that focuses more on explainability and morphologic description and we would have a supervised model that does crude diagnosis and the two would work essentially hand in hand. And for the supervised model, we would be able to generate a confidence score as to what the diagnosis is. Whereas for the morphologic description, we would not have that kind of a score.

The other aspect around the references is that there's quite a lot of research showing that you can somehow trace back what the model used in getting to a particular determination. Most of it is *post hoc* and is not based on what the model actually generated. So, there are a number of commercial tools, and I think ChatGPT, the latest version also now gives some references in instances where the information is retrieval-augmented generation, where it's directly pulled from the Internet. It's easy to do. Whereas with crude synthesis, it might be more difficult to generate what the origin might be. That's just based on how the feature space is organized. It might not be possible to map those abstract feature representations directly back to where the origin was. So, in my opinion, the current approaches for how you get to what the origin of those references would be are not holistic to be used in the healthcare space. We can get ideas or hints, but on the other hand, this is also true for a lot of the machine learning models, including those that have received FDA approval where interpretability techniques are giving hints and not crude information for what the abstract feature representation essentially is looking at.

So, I think in the short term, a combination of supervised models working with generative models for specific use cases like for pathology could be the solution in the

long run. We should keep an eye open for what's happening in the conventional machine learning space for enhancing interpretability and so forth. Thank you.

Dr. Bhatt: Great. Thank you so much. With the remaining few minutes, Dr. Maddox and then Dr. Jackson.

Dr. Maddox: Thank you. This is Thomas Maddox. Dr. Bhatia in particular. I'd just be interested in your thoughts specifically around operationalizing some of the user-output of generative model use cases. I think you make the very important point that humans-in-the-loop is a critical point of auditing. I think we've heard from you and others that we do need to think about making visible things like confidence scores, the cognitive, the cognition the machine is employing when it provides a particular diagnosis or recommendation. But I think we all know, and I'm certainly guilty of automation bias and I think it's probably unrealistic to expect every frontline clinician to pay attention to each of those metrics in the course of them going through their busy clinical day. So, to that end, what is the proper role for auditing? Are we envisioning that we want a periodic charge to some sort of auditor—informaticians, clinicians, some combination—that uses that information and ensures that we are dealing with information of sufficient confidence, sufficient cognition behind it. What are thoughts about that auditing process that allows us to think through that?

Dr. Bhatia: Thanks for your question. This is Parminder Bhatia. So, I think operationalizing these models, this is going to be key as we look across the spectrum of things. You touched upon areas of bringing human-in-the-loop. Initially, as we go into some of these use cases, it will be a lot of offline metrics. Some of the things which we talked about from preference side of things, getting extensive post-model monitoring of some of these things, so making sure it's validated at multiple sites to make sure it's kind of meeting the bar as well. So, I think that's going to be key to make sure they

meet to the standards as well. So it's going to require more benchmarking across multiple sites, kind of making sure it meets and actually meets the bar in a lot of those areas.

And then subsequently, I think in post-model deployment monitoring is going to become key. In some form, how do we capture that. And then have more periodic reviews of those. And I think that's where we need to have more machine learning and science behind that as well, like having mathematicians looking into that, providing inputs into that, where it'll help us to know not just-- Yes, there could be multiple ways in which we can get the output. It could be confidence score, it could be reasoning behind that, as well as multiple metrics as well, but not all of them might be relevant when the clinician is actually looking into the output, where the whole idea of these technologies is to reduce the overall cognitive overload by adding these additional components. We are in a way increasing that. So, I think a lot of this needs to move to offline and once we start meeting the accuracy as it is there with I think current models as well, we go with sensitivity/specificity, we are meeting 80, 90 and then we start to deploy those components as well. I think in similar form you'll look over here as well where initially you would want to do a lot of validation of these things, want to look into post-model monitoring and some of that will be discussed later today as well. And then kind of creating that flywheel going across, as well. And that's where in some form or capacity, it has to be a combination of AI and cloud capabilities as well, because you would need to create that flywheel that can go across multiple areas as well. Thanks.

Dr. Bhatt: Great. Thank you. For our panelists, I want the Committee to think-- Sorry, not panelists, the Committee to think a little bit about this as we move forward. The who, what, when, where, how of auditing, right? Just keep that in your head as

we're having more conversations. Let us end with Dr. Jackson as our final comment for this section.

Dr. Jackson: Thank you. This question is for any of the presenters today. One of the things we're tasked with thinking about risk management is controls for training. And I'm curious if you all have thought about "Is this going to be-- "Do people have to specialize in being able to use these models?" What are some of the training controls that will be needed for people who want to deploy these models in their hospital systems?

Dr. Rajpurkar: Pranav Rajpurkar. I think training's going to be important. We recently performed a study with 227 radiologists to find out what is the impact of AI assistance and we found that there was just incredible variability in the help that was produced. And what that goes to show is that there's something missing in the collaboration; recent studies now even showing that human plus AI performed worse than AI. All of this gets to the question of how do we enable that collaboration to succeed? And I think training's going to be a very important part of that. I think part of the problem, though, shouldn't just be on the users. The problem is often with the user interface and the developers and I think the onus in a lot of cases is if the impact of the tool isn't as great as we want it to be, then what do we need about the way in which we're interfacing that tool?

And regulatory wise, I think one of the things to think about is if we're thinking about separate regulations for different functionalities, for instance, if we want to have the ability to look at regions of the image which are being highlighted and have that same system be able to prioritize, are there existing regulations for that? And what do we need to do to be able to support the translation of systems that can create the sort of

trust and enable the sort of quick training to happen to really meet the user's needs and be able to improve their performance? So that's one way in which we think about it.

Dr. Bhatt: Thank you so much. It's now 12:02. We will now break for lunch. Thank you so much to our speakers this morning. For our audience who is here taking aggressive notes and to our Committee. Panel members, please do not discuss the meeting topic during lunch amongst yourselves or with any member of the audience. Panel members, please look at your meeting folder. It'll include information regarding your lunch arrangements. We will reconvene here at 1:00 p.m. to resume the Panel meeting. When we come back, it will be the Open Public Hearing portion of this Advisory Panel Meeting, so if you are registered to speak at the Open Public Hearing, please make sure that you're checked-in with the registration table just outside this meeting room. This meeting is now in recess for lunch. Thank you.

Open Public Hearing

Dr. Bhatt: Alright, 1:01 p.m., we can all start to trickle back in and get to our seats. Okay, welcome back everyone. This meeting is now reconvened. Before we proceed with the Open Public Hearing of this Advisory Panel, I just want to address our Digital Health Advisory Committee. You'll see in front of you three little post-it notes in order to, when we get to the question answer part, help us a little bit. If you have questions that arise as someone is speaking, feel free to write it on the note and pass it over to me after the speaker is done. That way if the questions are similar, we can compile them into one question and ask them to the speaker. And when those who speak come up afterwards, we'll just actually have them each come up one at a time and direct questions to them and hopefully that'll help us with our flow. So, thank you for cooperating with that.

Now, to 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 an individual's 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 a company 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 advise the Committee if you do not have any 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, James. 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 and their consideration of the issues before them. We ask that each presenter speak clearly to allow the transcriptionist to provide an accurate transcription of the proceedings of this meeting. The Panel appreciates that each speaker remains cognizant of their speaking time.

We have four requests to speak. The first one was to be a virtual presentation by Sophia Phillips. I don't know if we have that video. We do not. We will move on then. Our next presenter would be Annie Soh from the American Institute for Minimally Invasive Surgery. Is Annie Soh present today?

Okay. The next request is from Dr. Zimmerman, [Corrected 04:34:29 - VP of Translational Science] at Tempus AI. Dr. Zimmerman, are you here? Dr. Zimmerman, you have five minutes. Thank you so much for speaking with us.

Dr. Zimmerman: Is this for advancing? Alright. Good afternoon, everyone. And my thanks to the Committee and for the FDA for the opportunity to discuss this important topic. My name is Noah Zimmerman. I am a biomedical data scientist by training and the Vice-president of Translational Science at Tempus. At Tempus, our mission is to develop technology that allows every patient to benefit from the experiences of the patient that came before. To this end, we ingest and structure huge amounts of clinical data, integrate this data with data from our own clinical laboratories, and use it to create and deploy artificial intelligence-based tools that clinicians and researchers use to advance healthcare treatment and discovery. We see tremendous potential for the application of generative AI to the vast amount of data that exists within the US healthcare system. And I'm here today to talk about how we believe regulatory frameworks can accommodate innovation in generative AI and best serve clinicians and patients.

In a recently published survey of physicians, one out of five reported that they are already using ChatGPT for clinical tasks including documentation, differential diagnosis and even therapy selection. This rapid uptake of a new technology, especially for an industry that still loves its fax machines, is a testament to the tremendous benefits that frontline clinicians see in leveraging AI for patient care. And the potential benefits

are immense, including use cases currently under development such as medical note-taking, consultations and virtual second opinions and results interpretations for patients. I think the message here is that the cat is out of the bag. The train has left the station. The alternative is not a world where patients and providers lack access to generative AI. The reality is that they will turn to general purpose tools often without the appropriate guardrails in place. That's why it's so important for this Committee to convene today to catalyze a critical discussion about how to establish those guardrails, ensuring safety and effectiveness while fostering continued innovation. FDA has deep experience in premarket review of medical devices with different technologies and intended uses. What's clear is there's no one-size-fits-all approach to determine the evidence needed to support premarket review. Like so many advances that have come before, generative AI is just a technology that medical devices will use to achieve specific purposes. Focusing too much on the technology, likely misses the forest for the trees.

It's also important to note that not all generative AI used in healthcare is likely to be a medical device. For example, at Tempus we are exploring the use of large language models for the task of information extraction from a longitudinal patient record. We may ask the model to review the case and extract the stage of cancer at diagnosis or the results of a particular diagnostic test. Here the task is discreet; take information from a clinical document and the results are specific; a yes or a no, or a stage at diagnosis. Used in this context, the output is easily verifiable and evaluation can be conducted just like any other machine learning algorithm.

However, many other applications will likely be considered devices and when it is, the FDA should apply the same approach to authorize almost a thousand safe and effective AI enabled medical devices to date. FDA's current framework considers both the technology and its intended use. And FDA should continue to apply sound

regulatory and scientific judgment to evaluate these factors and consider specific benefits and risks when determining the appropriate information to support review.

All of this is not to say that there is nothing unique about generative AI. One particular challenge facing the medical AI research community is testing and evaluation of healthcare applications of generative AI. Unlike predictive AI, which works with defined boundaries, generative AI produces original content (text, sound, or images) that can appear highly human-like. It doesn't have the degree of obviousness that it is a computer generated, which may lead to an increased risk of automation bias. And so, we think that this introduces two risks that are particularly important with respect to generative AI; the risk that the output is inaccurate and the risk that the output is mistaken for human judgment.

Early evaluation efforts for GenAI have relied on well-defined tasks such as answering multiple choice questions from a medical board exam. These tasks have the benefit of being clearly defined with inputs and validated answers, making them easy to assess. However, they fall short in capturing the complexities of real-world use cases. In practice, healthcare applications of generative AI often operate in unbounded environments where the inputs are diverse, unpredictable, and context-dependent. For example, a model used in clinical decision support might need to interpret incomplete or ambiguous patient data or a chat bot for mental health support may need to respond empathetically while avoiding harmful advice. These real-world scenarios present unique challenges. How do we evaluate performance when the correct output isn't always clear? How do we ensure reliability when the same input might generate different outputs depending on subtle contextual factors?

These challenges underscore the need for more dynamic and real-world approaches to monitoring generative AI systems, especially as their behavior can evolve

over time and vary depending on context. This is where robust postmarket surveillance mechanisms, like a modernized MAUDE database can play a pivotal role in ensuring transparency, safety and continuous learning across the total product lifecycle. The medical device reporting database is a critical tool for postmarket surveillance, allowing manufacturers and the public to report adverse events associated with medical devices. However, as we think about the unique challenges of generative AI, we believe there's an opportunity to modernize MAUDE to better meet the needs of this evolving landscape.

Generative AI introduces risks that are more dynamic and context-dependent than traditional devices. For example, the nature of errors such as hallucinations or inaccuracies at output can vary depending on the input prompt, the user's interpretation or the context in which the tool has been used. Modernizing MAUDE could serve several important functions including increased transparency, crowdsourcing insights, and continuous learning.

Further, a robust postmarket database can be used as a resource for research and development, allowing manufacturers and academic institutions to analyze patterns and design safer, more effective iterations of their devices. In short, MAUDE has the potential to evolve into an active data-driven platform that supports transparency, postmarket surveillance and ongoing innovation. By embracing this vision, we can ensure that the entire ecosystem from regulators to developers to end users remain vigilant and adaptable in the face of rapidly changing technology landscape. Thank you.

Dr. Bhatt: Thanks very much. Our next speaker will be Dr. Bernardo Bizzo, Associate Chief Science Officer, the American College of Radiology, Data Science Institute.

Dr. Bizzo: Thank you. Good afternoon, everyone. Thank you very much for having me here today. It's a pleasure to be speaking on behalf of the American College of Radiology about our Assess-AI program. I'm a diagnostic radiologist by training, Associate Chief Science Officer at the college and Senior Director of the Mass General Brigham AI business. Those are my disclosures.

And just for a brief context, the ACR exchanges images and data with over 40,000 imaging entities across the country through its technology, the ACR Connect infrastructure for accreditation, registries, research, education, and very importantly today, AI programs. The Recognized Center for Healthcare-AI or ARCH-AI was launched recently and is the first national AI quality assurance program for radiology facilities with, at the moment, 18 recognized centers and almost 50 centers in the pipeline for engagement including international sites.

And as part of this Arch-AI umbrella, Assess-AI is a registry that provides real-world monitoring of imaging-based AI tools deployed in the clinical workflow. And Assess-AI allows for comparison of local performance metrics of these models to national benchmarks and also to facilities with similar characteristics such as region, facility type, trauma level, if it's urban or rural.

And it's currently being piloted with 15 sites with four AI vendors and platforms engaged for two FDA cleared clinical use cases—intracranial hemorrhage, ICH, or pulmonary embolism, PE on CT cases. And just at a high level, the current state, how it's being applied for AI monitoring. Each one of these sites or facilities are part of the program with existing ACR infrastructure can connect PACS data, including images and associated information, as well as medical records information and the AI manufacturer and platforms to ACR Connect and reach the ACR Assess environment where currently introduced two narrow AI use cases.

We collect site information about the final radiology report for each case that ran through the model, the AI model information, the manufacturer, the version of the software used and the output. This case-- Those are triage devices, just the binary classification of positive/negative for ICH or PE. And information about the case; the hospital, the facility, the patient demographics—age, sex, ethnicity,—the DICOM header information with technical details about the case that the model ran on.

And then at the ACR Data Science Institute, we are using large language models prompts to bias subject-matter experts to extract information from those recharge reports, in this case, if they are positive or negative for those two findings, if they are adequate or suboptimal in their quality as reported, and if the regionals had higher or low certainty when reporting those findings. And that data helps to generate analytics and reporting. So we can compare the report information with the model outputs and create dashboards with subgroup analysis. For example, assessing how AI impacts based on demographics, technical parameters, and also generate data regarding, for example, national incidence of findings.

And then, how this data can be distributed to different stakeholders at different levels. For example, for the site's local performance dashboards, benchmark reports for other sites that have similar characteristics I mentioned, and, of course, a national AI registry that can serve for much more information generation about these tools across the country. And that's the current state using narrow AI tools.

But of course, the future state is using GenAI use cases and draft reporting is the first one that we are thinking about going next. Instead of using the CNNs, of course, using visual language models to draft a radiology report. And then that information can be assessed using the same framework that is already established to assess draft report and final report differences.

And we're doing that through the Healthcare AI Challenge powered by the AI Arena. And you're going to hear more about that later today by Keith Dreyer. But in summary, it's a collaborative community for healthcare AI validation monitoring that includes the American College of Radiologists stakeholders, as well as Emory Healthcare, Mass General Brigham, University of Washington and Wisconsin.

So this data can be aggregated, healthcare professionals can be exposed to these AI tools and you can assess the local validation of these models as well as monitor their performance. With that, I want to be respectful of your time. Thank you very much and happy to answer any questions later.

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

Dr. Bhatt: Thank you so much, Dr. Bizzo. Will you stay there if you don't mind? That might make it easier. And for our Digital Health Advisory Committee, any questions for Dr. Bizzo? I'll start with one. First of all, excellent work and thank you for sharing how you think about it. As we're thinking about generative AI rather than machine learning per se, the way we've been using it with radiology, are there specific elements that you think about monitoring with regard to safety or efficacy that you would include in the kind of work that you're doing?

Dr. Bizzo: Yeah, sure. As you all know, this new technology acts in a probabilistic way. So, if you run the same data through the same model a hundred times, you may have different answers, hallucinations-- So we need to really focus on the content of the more output and how that, as was discussed previously by other presenters, can impact a patient safety and the healthcare providers not being biased if they're not experts in the area that they're focusing on. So, basically the content in the use case of drug reporting, I think there's a lot that we need to focus on with regards to the risks of the changing care management in a way that can affect patient safety and is an ongoing effort to think

about the specific variables that need to be accounted for. And that's the work we're doing at the college.

Dr. Bhatt: Are you able to share anything about those specific variables that you include to measure whether or not that is happening?

Dr. Bizzo: Sure.

Dr. Bhatt: Or if you're able to.

Dr. Bizzo: This is of course ongoing work. I think the number of changes compared to a final report when you look at the draft report in how clinically meaningful those changes are, because there are stylish changes, there are content changes. So, we are really focusing on how these tools can help with efficiency gains of the healthcare profession while maintaining patients' safety. So, if you look at the number of changes that could impact patient safety and patient care, if those are 1, 3, 5-- How much that would impact the management of that patient, how that impacts the efficiency of the healthcare professional using that tool. So, we are really starting to scratch the surface about the specific metrics as was well discussed in prior speakers' talks. But I think the number and the clinically relevance of those changes are the initial points.

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

Dr. Botsis: Taxiarchis Botsis. I think you more or less responded to my question that I'm about to ask. Are you also talking about specific metrics that should be used for specific use cases? And I think this is very much linked to Dr. Elkin's point before about having a card for each of those models with particular characteristics, but I want to really focus on the evaluation piece and that card.

Dr. Bizzo: Yeah. For sure, there's a lot to think about the specific metrics, and I think they're going to be use case, they are going to be model task-specific to a certain degree. The way we are envisioning—and Keith is going to talk more about the

Healthcare AI Challenge,—how different events, how this generative AI technology is applied to different types of data, generates different types of challenges. So, in a nutshell, I think we're going to need to expose those AI outputs to healthcare professionals at scale, to a lot of local validation, a lot of monitoring based on risk determination of how well that tool can impact the clinical care activity by those healthcare professionals.

Dr. Botsis: Taxiarchis Botsis. A follow up question on that. Do you specify certain thresholds for your use cases, thresholds that those models would exceed?

Dr. Bizzo: Yeah. So, we are thinking about threshold specifically about how well those models are performing. So, the initial thinking is having healthcare professionals rate the clinical skill level of, in this case, a draft report. How it was created by one or many of these generative AI tools. And then let's take the radiology example. Let's take chest X-ray. As a radiologist, I can look at a foundation model generated chest X-ray draft report and say, "This is at the level of a radiologist attending or at the level of a fellow." Good, but I need to make changes that change the clinical significance of this report, not just stylistic changes or at the level of a resident. It's okay, acceptable, but there is a lot to be changed to make sure that it's safe and effective. Or at the level of a student, where it's, with all respect, poor quality at that specific task—not at the level of an attending radiologist,—or it's unacceptable. So based on those rates of how well an expert professional judges the draft report of that use case, I can set thresholds where specific models that go below that threshold will not be exposed in subsequent steps of this challenge, of this event until we get to a point where a clinically safe threshold is defined for actual clinical use. Of course, bear in mind the whole regulatory aspect that needs to be taken into account before reaching that point.

Dr. Bhatt: Great. Thank you, Dr. Bizzo.

Dr. Bizzo: Sure. Thank you.

Dr. Bhatt: Does the Panel have any questions for Dr. Zimmerman? Oh, still for Dr. Bizzo. I'm sorry, I missed that. Dr. Bizzo, will you please come back? Dr. Elkin has a question.

Dr. Elkin: And you thought you were off the hook.

Dr. Bizzo: That's okay.

Dr. Elkin: So, one of the things with these that I always worried about, and I wondered how you're handling it, is that, first of all, there's trigger text that's used by many radiologic systems. So, you could say normal chest X-ray in a whole, your normal pops out and that's biasing in some way in the training data. But the thing that has been, I think, very difficult is that as people age, as they-- Different races and ethnicities may have different what looks like normal on, for example, a chest X-ray image. And how are you taking into account the breadth of normal? Because if you were to call pathology in all of the things that were normal, because it's very subtle in the way that it acts, it could be a burden to the practice. And I'm sure you're trying to improve efficiency. So, I'm wondering, how do you handle the routine variation in normal across the population?

Dr. Bizzo: Yes. Great question. And the same way that radiologists do it today. So, the actual images are being presented to the radiologists while they look at a draft report in this simulated virtual environment. And they can look at if those are generative changes, normal variations of anatomy, and how the generative AI draft report represents those changes. If called them pathologist disease findings that are worth mentioning on the report or as a radiologist that we say those are changes expected for the age of the patient. So of course, we need the expertise of the end-user to drive the significance of those findings and make sure that that's not biased in the results.

Dr. Bhatt: Thank you so much, Dr. Bizzo. Dr. Zimmerman, if you would rejoin us. Dr. Radman has a question.

Dr. Radman: Yes. Thomas Radman. Thank you. In your presentation you had these differences between generative AI and traditional machine learning, if you will, or algorithms. And I'm really interested to learn from your expertise in developing generative AI algorithms. So, I guess I just want to get a feel from your developer's perspective on the necessity for generative AI to have unbounded outputs, as you said. Or is it feasible to have more singular narrow scope or intended uses of a generative AI product? And I guess in particular, what is the cost-benefit trade-off? Because I could imagine an unbounded output could maybe-- If it's just a one-shot regulatory approval, you've got something that could impact a really broad range of conditions potentially. But perhaps it's not validated as thoroughly in some respect, and certainly in terms of development costs, in terms of a business case for going condition by condition versus unbounded product. And what are the trade-offs there?

Dr. Zimmerman: Yeah. It's a great question. So, I think to start, I think you hit on a really important point, which is, I think, sort of the canonical examples that most people in the room might think of with generative AI because of the way that it was rolled out in the form of a chatbot, have that unbounded nature. So, it's a natural language interaction. And so, as such, the unbounded part is you can write any English sentence in there and then it can produce any sentence as output. Of course, there are more narrowly defined cases as you sort of alluded to, one of which I think I brought up in my remarks, which was using it for very specific kinds of tasks. So, taking the context of a patient and asking about the stage of a disease or the presence or absence of a particular finding. And so, I do think that there are very clear ways that you can bound those environments. And I think really it's about the intended use of the device and how

you contextualize the GenAI. What is the component that the GenAI is doing and how does that relate to the intended use of the device?

Dr. Bhatt: Great, thank you so much. Dr. Soni.

Dr. Soni: Apurv Soni from UMass. The two things that you mentioned, cats out of the bag and trains leaving the platform, whichever one we want to stick with. Really curious to your opinion and considerations of how can we study that cat and train. And this is in the context of Commissioner Califf advocating for more pragmatic clinical trials. Director Tarver commented on real-world evidence. So, what would be some of the considerations where we can observe digital breadcrumbs or other things? Because if one in five providers, which may be an underestimate due to social desirability bias, are using some of these GenAI tools, what can we learn from it in vivo?

Dr. Zimmerman: Yeah, I mean, as I think that this is the opportunity for the FDA to operate as a central clearinghouse. So, I'll give an analogy, is the Adverse Event Reporting System for drugs. So, that system has existed for a long period of time. One of the things-- One of the many things that such a system enables is third parties to come in and study these things. So, there are entire academic labs and PhD students who have devoted their lives to mining that Adverse Event Reporting System. You understand things like toxicity profiles, drug-drug interactions, having a similar kind of resource for generative AI in practice would provide an enormous opportunity for the community to collectively study what it looks like in the real world. I think in the absence of that, and without the FDA's gravitas behind such an effort, what you'll find is local places that have this information and can get iteratively better. But in terms of pushing forward the entire community to create more safe and effective devices, I think it would be a real contribution.

Dr. Bhatt: When you were talking earlier and you mentioned "level the playing field" you were referring to also being able to report what doesn't go right, what are the errors, what happens. Can you talk a little bit more about error reporting and not in the academic sense, but in the industry and business-minded sense. Is that something that we can realistically ask of companies? What are your thoughts on that coming from the industry side and error reporting?

Dr. Zimmerman: Yeah. I think that postmarket surveillance is going to be a really critical part of this puzzle because as I mentioned in my opening-remarks, I think that the premarket presents unique challenges in evaluation. And so, I do think that industry will benefit from clear pathways towards how to get devices into market. And if that means additional guardrails in postmarketing surveillance, I think that that could be accommodated. I think there's ways we could do that, that the burden could be not super high on reporting. For example, you could imagine an API-like interface that a manufacturer could decide to build into such a device that would allow people to report unusual outcomes and sort of have that be part of the postmarket process.

Dr. Bhatt: Thank you Dr. Zimmerman. And just quickly, Dr. Bizzo, will you come back for one more question from Dr. Jackson? And that will conclude our question/answer session.

Dr. Jackson: Thank you. Not to put you on the hot seat, but you were talking about the data from over 40,000 entities that you're working with. And I'm curious because humans are fallible and clinicians are not excluded from that. Are there any controls in place to make sure that the images are real images, that they're not AI generated images? Because a lot of this is based on what is the input if it's an image. So, what are controls in place to make sure people are not creating images but using what was actually given to them?

Dr. Bizzo: Yes, of course. And just to clarify, the 40,000 facilities across the country are part of the ACR network. We have a subset of them engaged in the Assess-AI program that's currently being used for everyday cleared devices. So, only data from real clinical workflows where radiologists are looking at patient images, creating the report, and that information is being used for clinical care at those local facilities, are being transmitted to the ACR identified with all the proper guardrails regarding access to that data. And that is the data that's being used at the current state for the assessment with the AI results. So, you can see if those patients had or didn't have, based on the radiologist report, a finding of interest—intracranial hemorrhage or pulmonary embolism.—So there is no new data being created by those institutions, by other technologies, meaning AI generated data. But there could be, and I think it's a very valid question. In the current guardrails, since we're only using clinical data that's really being provided for patient care, that really simplifies that point. But we definitely need to keep that in mind for future endeavors with terms of AI and so forth.

Dr. Jackson: Can I ask a follow-up? So then I'm curious of your thoughts, thinking about scaling this. What would we need to think about? What are controls to think about so that people don't, right? Sometimes people misplace something, maybe they made a mistake because that happens and so they need to go back and correct even if they're using the image but they make a change to it. What are controls or guardrails to make sure that people are not manipulating the images that are the input that go in?

Dr. Bizzo: Yeah. We should leverage the existing controls in place in all those accredited facilities that are providing patient care and they are audited by the college, by other societies or other organizations, making sure that they are compliant with all the regulations. And we can add, on top of that, any additional guardrails if needed. But I do believe that those are very strict guidelines that already are in place so those

facilities can provide care for their patients. And we use secure connections to make sure that no data is interjected between that facility and the ACR. So yeah. I think relying on existing practice and make sure that they are enough for this new wave of technology of AI being used is certainly something that we keep in mind very, very critically.

Dr. Jackson: Thank you.

Dr. Bizzo: Sure.

Dr. Bhatt: Wonderful. Thank you both so much for your participation in this open session. So now, I now pronounce this session of the Open Public Hearing closed and it is now time for Panel deliberations and the discussion of FDA questions. The FDA has generated a series of questions for the Panel to consider. Ms. Aubrey Shick will present the questions and the Panel will deliberate amongst ourselves. Ms. Shick, please present the first question. After the Panel has discussed the question, I will summarize the recommendations for FDA before moving on to the next subpart of the question. Team, just in a less formal way, the goal is each of our questions that we received in advance have a subpart. What I'd like to do is achieve for the FDA three clear points of recommendation at a minimum for each subpart before we move on to the next one. So that's kind of our goal within this discussion. Oh, and they're in your folders as well. They might be in your email if you have your laptops open. I'm going to give you a minute to get those out. Okay. And Ms. Aubrey, you may present the first question.

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

Ms. Shick: Thank you, Dr. Bhatt. Our first question area is in regard to premarket performance evaluation. We would like the Committee to please discuss what specific information related to generative AI should be available to FDA to evaluate the safety and effectiveness of GenAI-enabled devices considering, for example, that foundation

models leveraged by the GenAI-enabled device will change over time and that there may be limited information available on the training data utilized for these pre-trained generative models.

So, there are four sub-questions. I will read the first one now. So, question 1.a in this premarket performance evaluation topic. What information should be included as part of a device's description or characterization in the premarket submission when the device is enabled by generative AI? For example, when a human is/is not intended to be in the loop, or if a device is intended only to recall information versus to generate new recommendations. What information is particularly valuable to evaluate the safety and effectiveness for devices enabled with generative AI in comparison to non-generative AI?

Dr. Bhatt: So maybe let's start with the first part of that question, which is, "What do we think should be included in the description or characterization in the device enabled by generative AI?" And if you have a present answer for one of the other parts, feel free as well. But just to help us organize.

Dr. Radman: Thomas Radman. I just had a clarifying question. So, I think the first part says, "Considering that GenAI devices can change over time," and I think we should distinguish that there's two different cases. There is a possibility as far as I'm concerned, that as far as I understand, GenAI devices can be locked. For example, we have GPT 3.5, GPT 4, etc. And then there's another separate question of continuous learning adaptive algorithms, which is a problem we need to figure out still for just pre-generative AI, right? And there's guidances on change control plans. So, I think it might help our deliberations if we separate those two questions like GenAI generally and then GenAI devices that continuously learn. Am I correct in that? GenAI does not have to continuously learn or is that a necessity?

Dr. Bhatt: I think that is very fair for us to break down as part of our response, which is lock GenAI versus allowing continuous learning. Is that okay? Alright. Dr. Maddox, you had a response?

Dr. Maddox: Yeah. So, I'll build on that. So just the three things that occurred to me and then I'd love people to edit or pile on is what's the use case or use cases envisioned by the particular device? Is it dynamic or not? And then, is it autonomous or not? Which speaks of the human-in-the-loop. And to me those are some of the things that I've heard to be important characteristics when it's time for the FDA to evaluate, its need to evaluate and the risk kind of category we put in place. But above reactions to that or building on that.

Dr. Bhatt: Tom, will you tell me the third of those intended use or use case dynamic or not, and-

Dr. Maddox: -Autonomous.-

Dr. Bhatt: -Autonomous. Thank you.

Dr. Stanley: If I can-- Oh. Can I build on that?

Dr. Elkin: Did you put me or her? Okay, so--

Dr. Bhatt: Why don't we go with-- Yes. You and then her. That's okay. Thank you.

Dr. Elkin: I'm sorry. So, I would like to build on Thomas's comments. I agree with them completely. I do think we also need to know the population on which the machine was trained and we need to know any results from prior testing, the methodology for the testing and the accuracy and the breadth of the testing. So, what populations was it tested on and what was the accuracy in those tests? Both and the safety of it, meaning that what were the error rates and qualitatively, although this is rarely reported, the things that-- Because they always answer, so the false positive rate and the confabulation rate are equivalent. So, the idea is that from those erroneous answers, how

much harm could come from those deviations from truth and that kind of an assessment would be lovely to have as you're making a decision about the safety profile of the AI and device. By the way, my large language model, when I put the question in, doesn't believe it's a device, it believes it's a software and it made a point of it to me. So, I thought I would give you that the LLM disagreed with me. I do believe it's a device. Thank you.

Dr. Bhatt: Dr. Stanley.

Dr. Stanley: Thank you. This is Laura Stanley. I was just going to quickly add on to that. I think that level of automation, that level of autonomy, I think perhaps needs a breakdown. We think about human-in-the-loop, human-out-of-the-loop, human-on-the-loop and how we classify that—if it's those three or if it's not,—I do think it's important that we know kind of where that human sits in that process and is it anticipated or is it not anticipated? If it is, how much? So, we can kind of think about those things of overreliance and vigilance, etc.

Dr. Bhatt: Dr. Botsis.

Dr. Botsis: Taxiarchis Botsis. So, to build a little bit on that about the use cases and how autonomous GenAI technology might be, I would think that it might be important really discuss the possibility of having a list of the particular functionalities or particular components that this particular GenAI supports for a certain use case because there may be a series of steps that one has to take there.

Dr. Bhatt: I have Dr. Soni next. Before that, I'm just going to, every couple comments, summarize what I'm writing so we're all on the same page. I have intended use or use case of the device with generative AI as well as the intended population for use is important to this Panel. I have the concept of "Is it dynamic or not?", "autonomous or not?", and importantly thinking about autonomy as human-in-the-loop,

human-out-of-the-loop, or not necessary, or on the loop. We'll come back to that in a little bit. And the anticipated human interaction that we expect. I have the definition of which model was trained, properties of the model and demographics of the training population from Dr. Elkin earlier in our conversation and understanding the functionalities and specifically the sensitivity or specificity of the device for those particular tests or functionalities. Combining comments from before with now. Dr. Soni.

Dr. Soni: In addition to who and what, I think we also need to consider where or the setting in which the medical device is being used, especially as a lot of our work and growing momentum is towards bringing healthcare to home. If a medical device that uses GenAI is going to be deployed in a decentralized setting, that needs to be pre-specified.

Dr. Bhatt: Dr. Shah next.

Dr. Shah: Right. Pratik Shah. So, I'm going to respond to the specific part A of what should be included. So, one thing that works for us is datasheets for datasheets, but before we accept a product in our lab to do any research, we have a standard pro forma datasheet where we have to fill out all the metrics of the data—the size of the data, the type of the data, the number of examples, where it came from, how long back it was collected.—So one recommendation could be that at the FDA we standardize this datasheet pro forma of accepting something to review, like, "Fill this datasheet out so we can track where all your data came from" and that could be just standardized completely. So there is no ambiguity when somebody submits something, what do they need to include? What do they not need to include? So that would really standardize the premarket. Just a characterization of what came in. That's one.

The second part is generative AI. I think a lot of people talked about it. When we do research on generative AI, there are different types of generative AI. So, if one

model converts one image to complete different image domain, that that model has never seen, that's a very unbound generative AI model. But if it converts an image that it has seen and generates some sort of notification or segmentation or some text report, that's kind of a different lower-bound of generative AI. So, it's very important that we have a clear distinguishing of what are the upper-bounds and lower-bounds of the models' generative capabilities on the tasks it is being asked to perform. Because you can have a lot of ambiguity in what the model is generating. Is it generating something it has already seen or it's generating something it has never seen? That's a second.

And the third is uncertainty estimations. So we use that all the time. We also need to know where the model is most uncertain or where your data was most uncertain that went into training the model. And there's Bayes by Backprop, Monte Carlo dropout, the technical ways to do it, but also the statistical and the common sense way of doing it is that asks counterfactual questions. Like what has this model not seen? What would it see that would break it, right? So that the uncertainty estimations are also very important deterministic things that we should really consider. So, we ask the tough questions upfront. What do you not know?

Dr. Bhatt: Dr. Kukafka?

Dr. Kukafka: Yeah. I think this might have been mentioned, but if a model is trained in a single institution and has a specific performance and it moves to another institution, usually the performance would certainly not be the same and most likely decline. So, whether that's part of the expectation that it would be trained in-- Depending on how it's being proposed to be deployed, I think that would be important to know.

Dr. Bhatt: Dr. Jackson next.

Dr. Jackson: I just wanted to add on more specifically to some things that were already shared. I think it'd be important to specify if they see it as a primary or secondary use.

Is it meant to support or is it meant to generate something brand new in the application? I also think it could be useful to indicate what is the percent of hallucinations or error-rate that they have found when they're submitting so that there's a general understanding of that when it is submitted.

Dr. Bhatt: Dr. Rariy, please.

Dr. Rariy: Yes, thank you. I think some of the other things that I'd like to consider is-- And some of it was touched on, but if we're looking at risk-mitigation-- So if the individuals who are submitting for FDA clearance or approval of the medical device already have a sense of what the limitations are of the model, then that needs to be specified or if there are specific requirements that are needed for the intended use to be executed. As an example, are professionals supposed to have specific training? Or end-users, do they need specific training to leverage the device in a particular way?

I also would recommend that we consider the cybersecurity and the data privacy aspect as well as part of that general standardized framework. So, what has gone into ensuring that there's not an ability to tamper with the product in current state or that there was that ability already in place when the pre-training was conducted. Similarly, the privacy standards that they applied for that data needs to be included. And then lastly, to include in this framework, perhaps the FDA can maybe come up with examples of what type of stress testing would be recommended so we can clearly determine, or at least articulate, the importance of stress testing to some extent that would need to happen for the device for submission.

Dr. Bhatt: We'll do Dr. Soni, then Dr. Elkin. Sorry, Dr. Radman first, Soni, Elkin. And we'll pause for a second to listen to the second part of this question because we're drifting a little into it, which is great. Dr. Radman, then Dr. Soni, then Dr. Elkin.

Dr. Radman: Thank you. Thomas Radman. And I apologize that pretty much all of our comments seem to be in multiple parts because it's such a broad topic. But I did want to reiterate the point Dr. Zimmerman actually brought up on the ability to build in a method to report for users to report any errors. And I'll add to his comment that ideally the particular report would go to a centralized independent database analogous to the FDA's Adverse Event database. Just any bias detection or hallucination, I think is a large part of this because of the challenges of doing an equally rigorous premarket evaluation as we currently are able to conduct with pre-generative AI devices.

And then I had this comment about-- We're talking about several Panel members correctly pointing out the need to describe the composition of the databases used, but we should recognize that the truly largest language model such as an open AI product, they currently do not disclose what's in their databases. And I'm sure there's industry representatives here that are interested in just directly using that foundation model in their products and perhaps that's using those products as a way of achieving equity where a smaller company can use these costly models. So, I think we should consider if we're talking about database description, are we talking about like an open AI or other very, very large language model and even are they able to describe their databases?

Dr. Bhatt: We'll do Dr. Soni, Dr. Elkin. Then to keep me honest, Diana Miller, I'd like you to come and just respond a little bit thinking about revealing foundation models in industry and error reporting in industry. And then we'll pick up with the other two. Dr. Soni.

Dr. Soni: I'll keep my comment brief because Dr. Radman covered a major part of it, which is I'm all for having nutrition labels for these models to know what's inside it and what we should expect from it. But the ingredients, if those ingredients are coming

from repurposing of foundation models, we also need to be practical about what is an onerous task that's unachievable versus what's for the safety. Recognizing that a lot of these tools are already starting to be used.

Dr. Bhatt: Dr. Elkin.

Dr. Elkin: So, you guys really made me think. And one of the things is that obviously these models can do a lot more than they're being tested to do. That would come before the FDA and I worry that there's variability in the accuracy of those ancillary tasks that have not been formally tested. And I think that this body, at some point, should discuss whether they would ban off-label use of the AI or as an alternative, some kind of postmarketing surveillance of off-label uses that would continue to monitor closely what they're doing because there is the chance of doing harm in situations like that. So, you guys are really bright. You're making me think a lot here and I think those are things that this Panel should at some point consider. The other piece of information that you'd have to gather from people premarket is any ways that they've used their technology that have failed miserably, crashed and burned, that definitely are unsafe. And I would like to see us have them report those upfront before approval so the indication can be accurate in terms of what it's used for.

Dr. Bhatt: Thank you. I'm going to ask everybody to look at the current question one more time. What information should be included as part of a device's description or characterization? For example, human in or not in the loop, device intention, whether it's a recall or new recommendations and safety and effectiveness. If you wouldn't mind, Ms. Shick, would you read part 1.b? Because I think we might be traveling between the two.

Ms. Shick: Thank you, Dr. Bhatt. So, we have 1.b up on the screen. And that question is what evidence specific to generative AI-enabled devices should the FDA

consider during premarket evaluation regarding performance evaluation and characteristics of the training data during the total product lifecycle to understand if a device is safe and effective?

Dr. Bhatt: Thank you. So, I think we've already begun answering this as well. Let me tell you where we are and then we will go on to more comments. I have broken us down into intended use, use case of the device with generative AI as well as indented population. We've also added the setting in which the AI is going to be used, whether that is a decentralized model or central, and primary versus secondary use. After use, we have the data itself—size, type, number of examples, standardized a pro forma datasheet perhaps to help us track our data.—We then have the model itself, the definition of which model was trained, the properties of said model, the demographics of the training population, where was it trained in terms of location in which it was trained, and does it include both a diverse dataset but also diverse location of care from where the data came? Does the model create bounded or unbounded results? And what are those upper and lower bounds? Requirements to leverage the device, logistic or training, and then cybersecurity and privacy precautions. I have those written out in longer sentences. I'm saving us time. We then talked about uncertainty estimations. Where your data or model is most uncertain, hallucination rates, sensitivity and specificity of the device for particular tasks or functionalities. We then addressed "dynamic or not," "autonomous or not," previous failures of the technology. That's where we are to date. Dr. Miller. I'm sorry, Ms. Miller.

Ms. Miller: Yes. Hello. You can hear me right? So, Diana Miller from the industry representative here. So, listening to all the comments from all of you, something comes to mind that the existing framework for approval devices from the FDA think covers a lot of these things. And I think a risk-based approach is feasible for this. And risk versus

benefit analysis should be developed when we submit this kind of medical devices and the risk mitigations apply to the intended use and the context of use are very important. So, it's very-- I wouldn't call it difficult, but I don't want to speculate about open-ended use cases. I think each use case has its value and each of these devices that deploy this technology, they should be reviewed in the context of use and in intended use. And with the risk/benefit in mind. And incorporating harm framework, like estimating harm and quantifying and special controls to address a harm is a framework that we already use today and we can leverage and apply that.

And going back to the risk/benefit, the comment that somebody made earlier that physicians are using this today, I just want to point out that we need to look forward and to be flexible in this new framework we developed to allow adoption because otherwise people will use it anyway and then the harm is worse. So, I'm not sure if I am clear enough, but since we don't have this kind of technology in medical devices today, and it's a technology that's coming to us, we should be a little flexible in deploying this framework at the beginning so we can all learn and climb in the same time. And the current framework with the risk assessments, it can work and we have cybersecurity assessments and all of those things that you talked about in place already.

Dr. Bhatt: Okay, thank you so much. So, two quick notes. Our question number two that we'll discuss later today will be on risk management. So, we'll come back to some of those thoughts there. As we go through the next couple comments, I want us to focus again on specifically generative AI as we give our answers for the sake of transcription in this discussion. And so a lot of our answers are becoming slightly more broad and moving towards AI in general. And we really need to help the FDA think specifically about generative AI. So, if you could please orient your comments towards that for the sake of transcription, that would be very helpful. Thank you.

Dr. Kukafka: Dr. Kukafka. So, I just wanted to push a little bit more on hallucinations because that was mentioned. And in the context of adverse events, which might-- I think it more traditionally are fairly transparent. I'm wondering if there's a need for defining how hallucinations are detected because they can be much less transparent.

Dr. Bhatt: Okay, thank you very much. Back to Dr. Stanley.

Dr. Stanley: Yes. This is Laura Stanley in hopes of helping to narrow in on this need. I don't think we've spoken a whole lot about the user interface design. I think we need to think about what that is and bringing in some fundamental human factor needs, and that surrounds ourselves around back to the user-trust safety. So, what does that user interface look like and what can it include to help that explainability need, that transparency need? Does the interface show regions that help explain the output? Giving that feedback to that user throughout will help gain trust? And ultimately, hopefully will gain trust and lead to safer outcomes. So just thinking about really what the interface needs to include. There's a lot of research in, more recently, in explainable AI, human AI that we could perhaps tap into.

Dr. Bhatt: Okay. Yes, exactly. So, Dr. Botsis, do you have comments on B? Because I feel like we've started to drift towards C. Okay.

Dr. Botsis: Yeah. I want to specifically refer to this sub-question here and maybe take a step back because we are talking about training here. And traditionally, training really requires some annotated and labeled datasets. And for GenAI, this process may be a little bit different. So, for example, when we try to train an LLM by doing some prompt engineering, there are certain steps and we have to understand what doesn't really go well in all those steps. We need more information about that and why the developers made the selection a specific selection to improve the performance through

prompt engineering. And I think this is an important part of the process. And what is the type of the training data that is used through that process?

Dr. Bhatt: Can I push on that just a little bit more? For the sake of this specific question then, what kind of evidence would get at the comment that you just made? What could we ask device manufacturers with GenAI to give us to help better understand that? And if you don't know now, we can come back to you.

Dr. Botsis: Like a suggestion that comes in mind right away is about the correction measures they took to improve the performance of the GenAI.

Dr. Bhatt: Thank you so much. Next we have--

Dr. Botsis: Oh, there's something else.

Dr. Bhatt: Dr. Soni.

Dr. Soni: Specifically commenting on the effectiveness part of GenAI. I think a lot of the predominating use cases for GenAI is to help improve the day-to-day workflow for providers in the setting where there is a mass exodus of healthcare providers. There is a lot of momentum and opportunity to use generative AI tools to improve their day-to-day workflow. So, one of the metrics need to be provider facing and how does that impact providers satisfaction or providers work to help make the decisions and make the clinical care more efficient and traditionally that has been seen as a preference or as a soft outcome, but maybe for generative AI, that may be one of the more opportunities to centralize on provider centered outcomes.

Dr. Bhatt: Okay. I'm going to take that comment. I'm just going to say something earlier today we also talked about the importance of having individuals who do a lot of kind of information science work and we do have the tap program through the FDA that puts people in touch with the right people to understand what you're doing with your information. I wonder if this is one of those upstream type programs that may be less of

a requirement and more of an opportunity. So, I'm just saying that to myself for the record. So, we have it for the transcript later.

Let's do Dr. Radman and then Dr. Rariy, and then let's look at C and see what we've covered for C.

Dr. Radman: Thomas Radman. I have a few comments and I'm happy to get cut off if I go too long or off topic, but someone mentioned restricting off-label use. The one thing that brings to mind is something I've heard frequently is that FDA does not regulate the practice of medicine. So, with that understanding, we should recognize that GenAI products have different risks than pre-GenAI products. For example, much more versatility to be used off-label than just a yes-no diagnostic that has a fixed input that we're used to how it's regulated. And then we've talked about breach of privacy, regurgitating someone's personal information, bias, misuse is part of the off-label hallucinations. So, if we agree that the risks of GenAI is higher than pre-generative AI products, now we're talking about FDA regulation where there's a substantial equivalence pathway. So, I think we maybe as a Panel should determine should the benefit of increase of accuracy have to be better than a pre-generative AI product if we are accepting that the risk is higher. It's not just the accuracy, if someone comes in and the accuracy is the same, but they're using GenAI, are we going to call that equivalent now accepting these risks. So that's one point.

And also, what I'm understanding from our panelists from the research community, that it's still kind of early days in the evaluation of these products and in terms of many aspects, determining bias, detecting hallucinations and just evaluation in general. I don't think that being early days should prevent someone from attempting to come on to impede progress, but we should recognize that that's also an increased risk that may justify at least a more rigorous evaluation process than a pre-generative AI

product would have. Then in terms of particularly speaking to evaluation, so because these devices have these nuances such as the way you engineer your prompt, how you ask generative AI, what you want could change your output and that could require different levels of training and also the outputs can vary as well because it can be conversational, if we're talking about a chatbot type of function that leads to broad interpretability by different users. So, I think the evaluation should be at an end-user level and there should be a stated diversity of the user base in terms of experience, demographics, like you name it, it should be very diverse and a sufficient number.

Dr. Bhatt: Great. Can we put up C? Because I think we've covered A and B well, and maybe Ms. Shick once you read C, I will let the group know what I think might answer, see what they've already said, we can add anything else to that segment. One thing I do want to mention is we're coming up against this quite a bit, and this is what is so challenging about generative AI. There are specifically the data, the model, the outputs that we are talking about, and those are all very kind of logistic parts that we can offer advice to the FDA. There is then this other side of the usability, who the user is, the flexibility in its use, the indications for use, which may not necessarily fall under our ability to enforce or state and yet is important to all of us. So, I am thinking about how we best convey the parts that may be easier to ask for require and the parts that are perhaps more suggestive about the practice of medicine, but very important to our group. And so maybe when you are telling me what you're thinking and I'm typing, feel free to say, here are some concrete logistic things that we have to have included and here are some concepts that are worrying me about the implementation of this in practice and that will just help me in my typing because as you know, I prefer to hand write. So, this typing thing is it's going okay. Ms. Shick, can we talk about C please?

Ms. Shick: Thank you, Dr. Bhatt. What new and unique risks related to usability may be introduced by generative AI compared to non-generative AI? What, if any, specific information relevant to healthcare professionals, patients and caregivers is needed to be conveyed to help improve transparency and/or to control these risks?

Dr. Bhatt: Wonderful. You see why I went here now? I think we've gone there. So, we have heard earlier from Dr. Stanley that user interface needs to help explain the output and give feedback throughout that product used to gain trust. We have also heard that it is important for individuals to recognize what the on-label indications are, therefore that there are off-label reasons that are not actually approved and be very clear on what those are. We've also spent time talking about training and the importance of understanding how things like changing your system prompt may actually change how you get an answer and that goes back to training recommendations that we may make to help improve both transparency, but help control risks. Are there other things that are new and unique risks related to usability that we have not yet talked about with regards specifically to generative AI that we should include in our response to this section? Dr. Rariy, I believe you were up next.

Dr. Rariy: Thank you. And I think this falls more in the prior around evidence, but one thing that I wanted us to consider is an opportunity for the device company to benchmark against existing AI devices or AI-enabled models that already exist and to look at if there's value added or if they're at parity or just some sort of looking around at the general device atmosphere. I think that would be helpful for information to come back with to the FDA on. The other thought is as we're looking at what is required in premarket-performance evaluation, I think it's always helpful to have a forward-thinking plan and so to include what could be your postmarketing strategy or your postmarketing plan, I think would be important specifically for generative AI where we

know that there could be subject to drift, there could be subject to other areas in information that is gathered from the real world. So I think that should be included in the premarketing evaluation.

Dr. Bhatt: Can I ask a quick clarifying question? An opportunity for the device to be benchmarked against other models, other devices with the same intended outcome and use that may include or may not include AI, that includes machine learning or that specifically includes GenAI, like theirs or all of the above.

Dr. Rariy: Probably all of the above, including if they're first to market, they may not have a specific comparative device, but perhaps they have the AI-generated models which they're currently leveraging and so there's that comparison, but at least an opportunity to think through what that comparison looks like.

Dr. Bhatt: Great. Dr. Jackson, and just let me know if you're on C or if you're thinking about A and B still that is okay for something additive.

Dr. Jackson: Thank you. I did have one additive for B and then I do have something for C. For B, and I may have missed this. I think it would be helpful to have information on the performance evaluation specific to different demographics and settings so they will have the overall performance, but how does that differ for the intended use, right? There's a difference in gender if there's a difference in if it's a rural clinic or something else. For example, foresee I think a new and unique risk, not just that the output cannot be reproduced, but as what was shared earlier that the reasoning will not, so they might still get the same answer, but the reasoning will not. The reasoning will not be reproduced and that should be clear that how they got to that answer may not be able to be regenerated. For what needs to be conveyed to help improve transparency, who is the intended user of the model? I think this goes to the training. Will it be techs? Will it be physicians, nurses, so that it's very clear who will need any training.

Dr. Bhatt: Dr. Shah.

Dr. Shah: Pratik Shah. I'm kind of B and C additive together kind of segueing. So, one is like it's very hard to have one regulatory framework for all generative models, so one recommendation for risk and usability would be for generative AI models that don't operate on images versus generative AI models that operate only on text, right? Because generative AI models that generate images do a very different risk and usability assessment than something that's just a chatbot that generates new text and then the hybrid models that kind of go sit in between both of them that use an image that's generated and then generate a text. So that's one usability aspect that's very important to point out. I'll give you a quick example, you could have a locked or a generative AI model that's not using chat GPT, but it's again a generative adversarial neural network that generates new images at the point of care. It's a generative model, but not a large language model or chat GPT that's still generated AI right? People still use that. So not everything is chat GPT or large language model. So that's one very important distinguishing factor to make for the regulatory framework to understand what we are reviewing and where it comes from and what the intended outcome is.

The C part which is related to it is one suggestion would be to actually include evidence to patients that this result was generated by an AI model eventually. I know that's a complex topic, but I feel like in the sake of transparency, if there is downstream clinical decision making happening from an AI-generated prompt, it may be good to label the initial AI-generated result that was generated. So, if it goes into the healthcare system and we lose track of it, we know where the AI-decision making was made. So, we can reverse engineer the path up and find out the node where the bifurcation happened. And that goes into the postmarket surveillance part and the trust and

transparency because if we don't tell people where AI intersected in their healthcare, we have no mechanism to track it back and fix it.

Dr. Bhatt: Dr. Elkin.

Dr. Elkin: Thank you. Peter Elkin, University of Buffalo. I remembered her instructions. So, I do have one for B, but Dr Botsis has made me think of this that when you're looking at getting a metric of whether the data was overtrained for a particular purpose, the number of epochs that it was trained on can be helpful. Normally, [Indiscernible - 05:48:26 (experts)] recommend you do about three to five epochs. I've started to see people who use 20 or a 100 epochs in order to train and in that case there's going to be high specificity for a particular task and a great fall off in functionality when the tasks differ from the one that's been overtrained. Then for the usability thing, I got to tell you a little story in order for my comment to make sense. In the years around the 2000, a bunch of us who were human factors engineering people went to NIST and we got them to recommend to you guys that every device be usability tested. And sure enough, the guidance came out from FDA that all devices need to be usability tested and they stopped there, no implementation guidance on what that meant.

So, most of the things were just, we showed it to some customers. There was no formal think-aloud method, Jakob Nielsen based a usability evaluation with a standard protocol in an unbiased way that took a number of participants and were watched unbiasedly for this. And there is a real protocol around usability testing that can be employed. I really urge the Committee to think about the last mile of this in putting in place a requirement for real usability testing before approval. Thank you.

Dr. Bhatt: Dr. Botsis.

Dr. Botsis: I'm not exactly sure whether this comment is so much related to one C. I think it is really important to appreciate the types of our output. So, for example, we

have a binary classification problem and we want to classify a case as A or B. In a non-generative AI approach, we would probably have a misclassification for a case A as B, but in the case of GenAI, you may have a third label unexpectedly. So, you really need to know that. The second example, when you want to summarize some long text, it is not a matter of whether you have a summary that may or may not fully represent the original text. It is, as we all know, like an output hallucination. You may have a summary that is completely different from what is described in the original text. I think that those matters are critical.

Dr. Maddox: Tom Maddox, WashU in St. Louis, I'm going to push back a little bit on Elkin because I'm thinking about this really important distinction that the FDA has about regulating devices but not regulating the practice of medicine. When we think about drugs and typical devices, the FDA appropriately stops at that point of how they approve it for a use case, they ensure the device your drug is safe and effective and then they give it to our clinician community to then use it as they see fit, recognizing that some things are off-label, some things are on-label. I'm trying to think of a similar framework here. I think what's important maybe is to stop at dictating use, I think that's sort of a fool's errand anyway, just given how heterogeneous this will be. But I actually think it's actually not appropriate for the FDA to be in that space just given how the statutes are set up.

But what does occur to me is one unique thing about generative AI is it very much interacts with the user in a way that's very different and changes the output in a way different than a pharmaceutical or a traditional device ever has been. So, it's made me think that interface, I'll call it the user interface between the GenAI device and its ultimate use might need some clarification by the FDA. The two things that I've heard, but just to outline them, are that we need to think about the user information that is

provided in a helpful way, I think Dr. Stanley's been very thoughtful on this, explainable AI, confidence intervals, etcetera. There may be some specifications there. And then, also user training, which Chevon and others have said, and we would need to think probably about that. I'll just leave it with this.

I think there's an analogy here from my world and NAMI's world of cardiology, and there is an antiarrhythmic medicine called Dofetilide, which has a lot of risks and because of that, the FDA has said, I must as a clinician prescribing, go through a prescribed training on how to use this med and I can't prescribe it without it. The industry I forget who makes the medicine, but they actually run the training so I had to go through a training before I could prescribe that med. So maybe there's a similar certification pathway to train me to interpret the user information to then equip me to use this in practice. So anyway, just a thought.

Dr. Bhatt: Perfect. I have Radman, then Clarkson, and then Shah. However, I want to give great credit to the FDA people who put these questions together. They appear to know what we're thinking about next before we think about it next. So, let's go ahead and show what question D or part D is, and then feel free to just tell me A, B, C, or D as you give your comments so that I can categorize. Ms. Schick.

Ms. Shick: Are there prospective performance metrics that are particularly suited or most informative for these technologies, given their complexity? What kinds of performance metrics are needed for multimodal systems for example, text image models where either inputs, outputs, or both could be multimodal? Performance metrics will typically vary with device intended use.

Some examples of known metrics to support discussion may be modality-specific such as generative text (perplexity, quantitative comparison to reference text) for generative images (Fréchet Inception Distance (FID), Structural Similarity Index,

Measure (SSIM) or for generative audio (Log-Spectral Distance, Perceptual Evaluation of Speech Quality) or may be functionally-based such as frequency and types of errors made by the generative AI-enabled device.

Dr. Bhatt: Okay, so we will go to-- You're still up, right? Yeah, Dr. Radman. And then just a reminder, A and B were particularly about evidence specific to GenAI devices premarket, what they should have shown us C is new and unique risks related to usability with these devices. And now D, really thinking about performance metrics that we would like to see.

Dr. Radman: Thank you. And giving credit to the FDA, I'll have to give them credit for taking this very robust and kind of varied conversation and disseminating into guidances and final regulations. So thank you. Thank you, Tom, you set up my points perfectly on labeling. I think that the labeling should be very clear that it should be in your face, this output, this chat advice. Especially when you're considering patients or lay users, has been generated by a chatbot, artificial intelligence chatbot. In terms of some of the training points that Tom was getting at. So the repeatability or the temperature setting of the device and its propensity to give different outputs for the same input is important part of the training for a user, but it should also be clear the input, the premarket information to the FDA should also have some-- And this is already included for traditional algorithm review, like a repeatability and reproducibility that really hasn't been spoken about yet. But how consistent is it with the same inputs on a large scale.

We should recognize that some of the latest literature shows that repeating the same inputs is now a mechanism to recognize what part of a response is hallucination. Usually the varied part if you're repeating, and then the part that stays constant is the truth of the response. There's also literature that shows that repeating prompts actually

increases accuracy so if you take a generative AI for a task and you have it answer something once versus answering something three times, it does better in the three-time case. So that should be part of the premarket information, how it's going to be used, education about repeatability, what's the temperature, what are the guardrails, what's the complexity of the model that's directly linked to its propensity to hallucinate. So the larger the model, the more a premarket review should be looking for concerns on hallucinations.

So that's maybe D and getting it to-- Sorry, that was C and maybe getting it to B. I also wanted to mention, I think this is either B or D, we talked about broad-use case generative AI or unbounded, and Dr. Mahmood gave the example earlier that he had this ability to classify now 108 cancers, that sounds so amazing, but we should recognize that some of those cancers are very rare cancers. So if you consider a device like that and you just have a broad indication of it can detect these 108 cancers, but one of those cancers occurred 1 in a 1000 just making up a number, we didn't really validate that one cancer. So we should, even though it is a broad use, we should be breaking it down and particularly paying attention to the rarest cases, those are probably the hardest to classify.

That same point goes for users when you have broad demographics using a device, whether that be clinicians or patients. I'll give the example of language, the rarest language used in the population has been documented to perform the worst in generative AI use. So I think a focus on the rarest class is sort of like this idea of asymptotically arguments. If you're thinking a philosophical construct to prove something, take the hardest, most extreme case and that would be in this general of AI analogy, the rarest classification or user group. Then I just wanted to mention there's a publication many of us might be aware of it, co-authored by actually someone who used

to be at the FDA Digital Health Center for Excellence, Bakul Patel, where he talks about regulating generative AI as doctors. It's unbounded, doctors are also unbounded. So how do we regulate them? They have to take a test every year and there's challenges with giving something that's ingesting everything in the web, constantly on testing it. But I think that's just a really interesting idea for this Panel to consider, or at least I'm just stating that idea for the record.

Dr. Bhatt: Dr. Clarkson, and then Dr. Shah.

Dr. Clarkson: Hi Melissa Clarkson, consumer representative. I want to follow up on Dr. Elkins comment about, and I'm back on C, I believe it was talking about risks. It mentioned patients and caregivers, and I think that was a reminder that we've mostly been talking about in-clinical use of generative AI, but this is going to end up in homes or it might be assigned in a clinic and then you bring it home. I mean, this is going to be used by people who do not have clinical training and I think it is at the fuzzy boundary of-- Is this really within the FDA's boundaries? But studies to understand how non-clinical people seeing that generative-AI output understand it? How they react to it? How do they interpret its authority? Because it could be like if I go to a very prestigious medical center, I might say, oh, this physician has much more authority than my little hometown doc. How do they perceive the authority of generative AI? Especially, if it's using my name and it seems very personalized and it knows all about me. Do I understand that it's even more authoritative as my doctor I haven't seen in three years because it's too expensive. I really have no other access to healthcare. So I think some sort of study going in that direction of how do non-clinician viewing this output understand the output.

Dr. Bhatt: We'll go to Dr. Shah and Dr. Soni. I just want to make one comment. Thank you Dr. Clarkson for that. There was a day about a year ago where I called a

chatbot. Really nice. I was like, wow, that's a really nice chatbot. And then I just thought about what I did and so I love the comments you just made. Okay, over to Dr. Shah and then Dr. Soni.

Dr. Shah: Okay, Pratik Shah. I'm answering specifically 1D and great technical questions. I really like it. Thank you for the FDA team for putting it up. So I think this is a specific example where a generative model has created a new image that potentially did not exist and you're trying to characterize the quality of that image, and you're using Pearson correlation coefficient, SSIM, etcetera. And we have dealt with this problem. And I think my recommendation in general for the regulatory science perspective of this is that we need to do all of it. We need to have a really good computer vision validation metrics, which has no human involved. So it could be pixel by pixel quality assessment of the generated images by Pearson correlation coefficient, SSIM, PSNR and once the image is signed off by those computational metrics, the other aspect to consider would be would a clinician sign off on that image?

And we have done that work because we have to do a common-sense approach that once we have the generative AI model image validated, we then have to go to the clinician and say, okay, will you sign off on this image? And what would be the standard of care you would give using this image? Right? So that's the second sub-part, which really needs to be considered. The errors that will compound could come from the computer vision problem where the Pearson correlation coefficient or SSIM wasn't good enough, and that compounded in the clinical decision making, or it could be the clinical decision making was independently errored in its own errors, which had nothing to do with the image, right? Because clinicians sometimes don't agree on rare subtypes. So that's the two part, and then the very important part is how do you communicate it to people and understanding the multimodal part.

So I think for the multimodal part, the communication and the interpretation of this model is very important. And I'll give you a very quick example: People confuse interpretability with explainability in machine learning models and AI models. Explainability means saliency maps or some sort of grad-CAM where you can show what the model was looking at when it got activated, but it doesn't have much to do with why it chose that activation path to make the decision. So I think it's very important to kind of communicate that explainability of the model may be unlinked to interpretability and saliency maps or activation maps that generative models use may not have much to do with why the decision. So the solution to that problem is potentially looking at the saliency activation maps and then teasing out specific features from those saliency activation maps, which actually led to the deterministic decision made by the model. Because if we don't have that link to the output, and just an activation, we are confounded in the technical side.

Dr. Bhatt: Okay. We are slowly coming up on time, so we have Dr. Soni and Dr. Maddox. To the FDA team, I will ask a question which is we are going in order of the agenda given today. I'm curious as to whether the break can be given before my summary to Mr. Tazbaz. Is that allowed or possible? Great, thank you very much. I will say probably for the sake of the FDA, nobody can talk to me during the break. I will just sit here by myself with my computer so that we don't have anything that was not recorded for public record.

Mr. Tazbaz: So, the question, I guess, is are we going to have an opportunity to ask clarifying questions?

Dr. Bhatt: Yes, clarifying questions first would be great. So, I will do a version of a summary, clarifying questions and then a final summary.

Mr. Tazbaz: Okay, wonderful. Thank you.

Dr. Bhatt: To Dr. Soni please

Dr. Soni: A comment on 1D. Those metrics are really important and I consider them as safety metrics because that helps understand and continue to prospectively evaluate whether or not the models continue to operate as intended. Very challenging with generative AI because there is variable input and potentially variable output or by design variable output. But I do also want to highlight that, and this may vary based on intended-use case, but there may be a need to not lose track of some of the traditional metrics which we have learned to understand a lot of devices and decision-making tools which is sensitivity, specificity, positive predictive value, negative predictive value, precision and some of those metrics still need to be part of the discussion because if metrics are only understandable by very highly-specialized workforce, those become challenging. Again, this may become a concept part, but I think the safety and effectiveness we need to maybe think about them as separate metrics.

Dr. Bhatt: Thank you. We'll do Dr. Maddox then Dr. Elkin.

Dr. Maddox: So to continue my role as a contrarian, Dr. Shah, you talked about evaluating the use in this case of an image example. Why is that not getting into the practice of medicine?

Dr. Shah: Sorry, repeat your question again.

Dr. Maddox: So we can't regulate the use of the practice of medicine.

Dr. Shah: Correct.

Dr. Maddox: And if I understood you right, maybe I didn't, was that you said one thing we need to think about is once an image is generated that we then need to evaluate how the clinician is using that and that struck me as a little bit out of bounds here, but maybe I misunderstood.

Dr. Shah: Right, so I think different settings require different oversight, I assume. So my specific comment was that if you have an image that was generated to train a deep-learning model and the image was then used by clinicians, validated by clinicians to help the deep-learning model perform better or performance metrics, that's a valid oversight of the model. That was my point.

Dr. Maddox: Thank you.

Dr. Bhatt: Dr. Elkin, next.

Dr. Elkin: Thank you. I've been thinking a lot about all the comments and there's some really wonderful things to build off from Dr. Maddox's comment. There's a difference between drug companies and the companies that are marketing the AI in terms of their usual practices. I think that in order to above all else do no harm, we really want them to have good guidance from the FDA about how they can market these things. I think it falls short of regulating practice, but it does give guidance to the people so that it's not a slippery slope to this model's approved for one thing, but it can do everything and because these are very broad and interesting, it made me think of an example that might help your case, Dr. Shah, which is let's say you have an MRI and you take from the MRI and generate an image, an augmented reality for the neurosurgeon taking out a brain tumor. You could see that as a generative image that would come out of one of these things that would've real effect on patients. Just because it was proven in that circumstance doesn't mean it's going to help a cardiac surgeon find all the calcifications or a colorectal surgeon find the extent of the tumor, right? So we really want to give good guidance to how these things can be marketed.

Dr. Shah: So I think to quickly, so that's why I suggested labeling where it came from and what it was intended to do and if you do something else with it, then you are

taking the risk, not the FDA, not the people who made the model because you decided to use a drug for an off-label recommendation.

Dr. Bhatt: Okay, with that lovely spirited discussion, we're going to move for a minute now to give a short summary and then ask the FDA if they have clarifying questions for the group. So those who didn't get to speak yet another time, I think there's still opportunity for discussion. So I'm going to go back to the beginning with A. When we are looking at this, we are looking at-- I'm going to do the short form, tell me if they want longer. Intended use that includes use of the device use cases intended population, where it's going to be used, primary versus secondary use. We're talking about data, we would like to make sure that the FDA recognizes we need the size, the type, the number of examples. We need to think about standardized proforma data sheets that include all of this information and allow for tracking.

We do need uncertainty estimations where your data model is most uncertain and how it was trained and whether it was overtrained. For the model itself, we have the majority of information that we are providing to the FDA about our opinions. This includes a definition of training of the model properties, demographics of the population, and does it include a diverse dataset but also location of care. Importantly, bounded or unbounded is unique to generative AI and important part of what we would like people to be able to report in the premarket state. Cybersecurity and privacy standards as with other AI. However, recognizing that these are models that may change over time, so will have unique needs that may not otherwise exist when these algorithms are changed over time. Then for B specifically to generative AI, we do think that there might be an opportunity for the device to state how it benchmarks against other models that may have existed previously without generative AI or other generative-AI models similar to it. We do think that it would be nice for them to include a little bit about

postmarketing plan in their premarketing conversation, but we will get to postmarket surveillance tomorrow. We would like information on the performance in different populations and settings as well as information on repeatability and reproducibility that is important and unique to generative AI as if they have information on how different temperatures offer different responses that would be interesting to those who are going to either use the device or build off of it.

Also important to recognize how narrow the indication for which it was created is how many epochs were actually used or evaluated. Hallucination rate, we want to understand sensitivity and specificity of the device for particular tasks and hallucination rates for those tests and functionalities. And then lastly, we have not had a thorough conversation yet, but we will about the breakdown of autonomy, which also goes to the degree to which it affects direct patient care, bringing in what Dr. Elkin just ended with talking about neurosurgery versus cardiology.

Lastly, for C specific to usability, we are thinking about the user interface and it's explaining the output and giving feedback throughout to gain trust. We are thinking about the fact that the output and reasoning may not be reproduced as with prior technologies and therefore that should be something that users are aware of. We recognize that generative AI models may operate on images versus text versus multimodal, and we likely have specific metrics that are coming in D depending on what they are actually operating on.

We do want to talk about requirements to leverage the device logistics and training, and we may recognize that outputs of generative AI may not be what clinicians are used to of A or B, but may create a novel response, again, necessitating section D, the metrics for those novel responses. Lastly, we commented on C how we need studies to understand how non-clinical people understand these outputs because the location of

use will also likely influence who is using this for direct healthcare, and that is not always a super-trained clinician.

Lastly, for D four metrics, we agree with the known metrics for safety that you have suggested here. We do think we need to continue to evaluate models and models remain accurate without drift over time and we need a metric to be able to do that. Then finally, we do think that it's important that these companies demonstrate good computer vision metrics without human input first. Then what does the human input of clinical evaluation say? And then lastly, how do they clearly communicate those outputs. FDA, any clarifying questions of what our group has discussed?

Mr. Tazbaz: Well, thank you Dr. Bhatt. Give it up. Thank you for such a rich discussion. I know this is a very complex topic and it is very engaging and we're actually having a hard time not to engage as we're instructed to do. But we do have a few quick hit clarifying questions and maybe three additional questions that probably are not going to be answered here, but it's something that we should definitely take back and address at a later time. I'll maybe just go through the quick hit questions and one is we've heard the term dynamic AI models being used a few times and we're trying to understand what that is mean because we're trying to align on terminology as well. Can't remember who used that, but there's been the term dynamic AI, and I want to make an assumption, which I don't really want to do, but is that adaptive AI? Maybe? Okay.

Dr. Maddox: I used it and you're right, that's a more precise term, but I used it initially and that's a more precise term, but I also would look for validation on that. Is that consistent with what others are thinking? Okay.

Dr. Radman: Yeah, I think one of my first comments was about continuously learning there and then that brought us into the dynamic.

Mr. Tazbaz: Okay, thank you. Thank you. Yeah. Okay, so adaptive artificial intelligence. Okay, the next one, hopefully Dr. Rariy, you mentioned about benchmarking against other models. Where should those benchmarks and performance metrics should be stored? I guess who's responsible for maintaining those benchmarks?

Dr. Rariy: I think it's a good question. The only thing I can lean on is where historically medical device benchmarks have been stored. So perhaps looking at the literature as an example. That's where I would start, but it might need to adapt over time because while we've seen quite a bit of literature in the generative AI space that may or may not be representative of the devices that we would see through the FDA.

Mr. Tazbaz: Wonderful, thank you. So here are some questions that are very generative AI specific that we captured from the discussions and these are really important topics that we do need to address. And the first one is around hallucination and drift detection. Dr. Rariy you mentioned about stress testing models and coming up with methodologies and perhaps tools to be able to stress test. From a regulation perspective, should the requirement for breaking these models and anyone who's done software development understands ethical hacking, right? You have ethical hackers in your staff that tries to break these applications, should that be a requirement for generative AI where there is a stress testing requirement for regulatory approvals that allows you to understand the breaking points when something is going to hallucinate versus drift? I would love to hear some comments on that.

Dr. Rariy: I have an additional comment there I suggest. Yes, and perhaps we could base it based on risk. So one of the comments that I was going to propose is specifically the metric would be an error-rate analysis. So not only describing the type of error, the frequency of the error, but the severity of the error as well. So this comes back to the stress testing requirements can be based on risk of the medical device, risk of harm if

you will. Is the risk high enough to cause death or is the risk high enough where there's misinformation and there's different stress testing recommendations dependent on the risk.

Mr. Tazbaz: Okay, so the requirement for stress testing or the level of stress test needs to be aligned to the risk associated to the model.

Dr. Rariy: That would be my recommendation, but I would certainly want to hear others.

Mr. Tazbaz: Thank you.

Dr. Jackson: I would say that the ownness of responsibility should be on the company. I do think that it does need to happen, the stress testing and I think about, I know a lot of our presentations today we're talking about radiology and different things, but coming from behavioral medicine and mental health perspective, we've seen that with chatbots. I think sometimes we conflate generative AI and we forget generative AI in healthcare needs to be vastly different than just thinking about how we produce GenAI in general. So I think companies do need to show along the same line of responsibility for clinicians of do no harm, how far it would take or how long it would take for hallucination so that people can mitigate the risk of harm to patients that they're referring to or that they're using the GenAI for.

Mr. Tazbaz: Okay. And which is going back to Dr Bhatt's question around the industry's responsibility and some of these performance management.

Dr. Elkin: So I think it's absolutely necessary and I'll give you a couple of good examples of why and then I've got another recommendation aligned with that. So as you're debugging these, as you put them up and debug them, there's some unpredictability in the results that you get and the wider the audience that you can stress test it on, the better you'll be able to adjust the parameterization of the models because

in addition to the training of the models, you get to parameterize them if you really want, I can do a teaching session on it, but the idea is that there's a good lot of choices that you can make. They have real impact in the confabulation rates and so forth, but the more people you get to try it and you tell them to try to break it, the more likely you find a good share of these upfront before you would release it in any kind of way.

We do this routinely, and I tell you in my department, people have real fun trying to break the models I build because they love to tell me if they did it because they're real smart and it shows they're smart and so they have fun. It's actually a fun activity for the companies. I'm sure they would enjoy it and it would help them to hone the parameters that they put in. The other thing you didn't say, but I think might be aligned with that is maybe people should report Shapley curves for the models. These are deep-learning models and they're not explainable completely, but we can back propagate through it with Shapley curves and try to get a sense of what the most important features are. I know a lot of people care a lot about explainable AI, I would love to see that. Of course I was just in San Francisco and the Waymo were there, self-driving cars and nobody who got into one of those cabs actually was able to ask ahead of time how the model was trained in the car that allowed them to drive through San Francisco safely.

They were happy that they had a good safety profile and they gave their credit card and they took them where they wanted to go and they got in and out and even people I know are very sensible and careful people. Matter of fact, you guys know David Bates, he was one of the people who did this. He got in one of these things and he's a safety person and he couldn't know how these things were trained, but he had a good assurance that safety profile was good and he was willing to trust that. That's sort

of interesting for someone who's so well-known in patient safety, right. Anyway, thought I'd share.

Mr. Tazbaz: And by the way, it is a good example of using different industries who are applying autonomous use to different industry use cases and can we learn from the guardrails that they've established and something that we have been asking. Just to be respectful of time, this one may not necessarily be something that we're going to answer today, but there was a discussion on third party foundational model usage in applications and from our point of view, we do wonder about what information should be actually available to us during submissions if a third-party foundational model was used as the underlying model for applications to be applied to a specific kind of healthcare use case. We've been calling it either multi-tiered or multi-layered architecture, so that's in an area that we are very interested in kind of pursuing additional information. That said, I think the last few comments kind of goes back into or ties directly into the question that we have around usability.

So it's about the interaction with a device, I think is what we're talking about, the usability. The question for us is how should we thinking about performance management and the guardrails that should be applied to usability considerations around how someone is interacting and whether the performance changes based on the interaction with a specific device.

I don't think we talk about this often, but in the early days there were a lot of actually psychologists looking at what people were doing and part of it was that they wanted to break the model and so they were trying to see what that break point was and that's a behavioral pattern in engaging in. So the question now for all of us to consider here is that in medical devices, how should we be thinking about this thing from an

engineering point of view of how the user is interacting and whether performance changes based on that interaction? Would love to hear any comments on that.

Dr. Bhatt: So the group at NYU is doing some really good work in this and entering healthcare, which is, it's almost like the behavioral, I don't want to call it economics, but the behavioral manipulation with AI that happens. Right? Which is they're in fact studying if you are using the device, does it make you better at it? We talked about this earlier, right? Sometimes AI plus human is worse, but also what is it telling you about yourself, about your knowledge level, about your confidence in decision making? How does it affect your ability to care for that patient if something tells you something different than what you had? I think it's a growing field, but it is a young field and so I too would be interested in what the group says, but I would also be careful. I think there'll be a lot more data about this and this field is about to grow and luckily we'll be together for three years, but I think this field is about to grow.

Mr. Tazbaz: And I guess a more concrete question on this one is should that be part of the evaluation criteria.

Dr. Bhatt: As far as performance metrics that may influence outcomes for patients? It goes back to what Dr. Rariy was saying. I think a lot of our conversation the group can agree or disagree entails what exactly are you doing for the patient, how much harm could you possibly cause by doing this thing using generative AI? And I think when we base it on that, are we triaging? Are we giving an exact diagnosis, are we affecting the operating room? I think there are different categories and based on that there is probably a different level of stress testing. I would say how a human interacts with it and changes their clinical care probably falls in that realm. If you're going to change something that's not going to really affect long-term care or short-term acute care, that might be different than acute difference in what you would do based on interaction, I'd

almost maybe think about including that in stress testing. Dr. Maddox, you can be contrary if you want.

Dr. Maddox: No, I agree with that. Shoot, I wish I could. I think that's right and it also makes me think maybe this is a role for a calibrating temperature to risk and so the more risky it is, maybe it's more deterministic, which in turn will lead it less dependent on the user interaction with it about the output. So this may be something that you evaluate on a spectrum along with how you proposed it on risk.

Dr. Radman: This is a good time to bring up a point that we haven't discussed is that the support provided by the provider of the device, the GenAI, is part of the usability. So I'm just thinking of a study where they were using online smoking cessation, which is not GenAI but it's a digital health product and they had one group just using it. One group was given broadband access and one group was given broadband access and access to a digital health literacy coach. The group with the coaching had almost double the treatment response. So just like simple customer support provided by a company could really increase the efficacy of a device and so that's just a concept we might want to consider too.

Ms. Miller: Thank you. I think I have comments through a lot of things that I didn't get to talk to. So there might be a little bit in the random order, but I'm going to start from-- Because how you summarize, I think, structured the way I wanted to talk about it. So, going back to Dr. Thomas. When manufacturing, we are very careful and mindful to serve the patient when deploying this kind of technology and our doer is best approach. So to your point, we don't use the technology just because the technology is available. So the fact we look at the problem and I try to find the best solution to the problem, and I think that still applies for GenAI, they probably be different way to solve

the same problem and not using GenAI. With that being said, the risk that the device entails also targets in a way how we pick the best solution to it.

So in the same fashion, the stress-based testing that we are talking about should be targeted based on risk. I'm worried of enforcing a stress-based test for all devices and then they're going to be companies trying to pop up, "Okay, give it to me, I'll test it for you," and that kind of stuff. So, I want to be mindful of that and when we do need to enforce a stress requirement in a device. Now going back and what the benchmarking would be, and this is a new technology and that kind of stuff, so it might just hinder deploying our devices if we go just blindly. On a matter of usability and transparency, when we submit devices right now and we talk about this in PCCP and in our regulatory frameworks, we also include the targeted-intended population and to what you mentioned earlier, the intended population for which a device is intended to be used as part of the device package in a way.

So therefore there are different usability considerations and different usability studies that the manufacturer does when they are targeting a population. And we will need to be mindful of it. I don't know how to explain this, but I'm going to maybe use a little different words or too much confession. For example, for the sake of transferability and usability, you might disclose too much information when it becomes harmful because it's very confusing. So I call this a confession disease because you and I apologize, I want to make my point clear. It has to be targeted to the population you're intended to target for because you have a different kind of transparency for a patient and different kind of transparency to a physician and different, and you mentioned very greatly that practice of medicine varies and in cities versus its undeserved places, the practice of medicine varies. So depending on what your target population is you need to design it differently. And with that I think also the benefit of how do you design this

device to allow adoption to healthcare and underserved population, may outweigh some of these risks that we don't exactly know today how to design force. And I had one more comment that I'm trying to remember. I gather all of them when you guys are talking. No, I think the usability was the one. Thank you.

Dr. Soni: I'll just say in relation to usability testing, we do have some opportunities with use of simulation labs that have been more widely used for learning, but in many ways there is a great need for learning and a very steeper learning curve on how we are using and how the end-user is using GenAI. So usability testing data that comes from simulation environments where it's more contrived-use cases and we're able to get at specific components of GenAI and its implications on interactions and decision-making may be helpful and dovetailing that with where some of the benchmarks should sit. I think the idea that was brought up by Tempest spokesperson of having clearinghouses where we can continue to be able to look at prospective validation as well as ongoing monitoring is helpful to be able to continue to learn from ongoing practice and use of it. I'm not sure how much clearinghouse in that concept fits into the roadmap for the data that gets generated for its evaluation of GenAI, but that may be an important consideration.

Dr. Jackson: To your question about usability influencing performance metrics. I believe that it does, especially, I'm just thinking from mental health examples, there's research that shows if it's a person that end-user is a patient who is using the generative AI, the relationship feels different and they see it more as an authority to Dr. Clarkson's point compared to if it's a clinician who's using it and how they're thinking about it. So I think then the perceived performance could be different even how a patient may say this is being more effective than it actually is because of how they're feeling about it. The relationship compared to a clinician and the outcomes that they may get. I also

think when you're saying should it be considered, I think one of the things we haven't thought about is consideration to what weight. There's lots of things to be considered, but it is something that is a 90% weight in the consideration. Is it a 10% weight? So I think that is something also to be considered when evaluating how much weight do some of the considerations should have or do have.

Dr. Bhatt: I think Radman, then Miller and then Shah.

Dr. Radman: Thomas, I just want to make a comment on the question on ethical or white hat hacking. I think that it's actually very important for the reason that there's this concern of the ability to generate AI models to ingest any tests or benchmark that we might be able to put forth. So once a benchmark becomes a target, it's no longer effective. It was mentioned, actually the first comment today, this idea of these quality assurance labs, or CHAI, and I think Dr. Bizzo's organization might be related to this assurance function, if I'm understanding correctly-- ARCH-AI. And FDA has some familiarity in they already do a third-party review for some products, so passing off some review functions to an independent source and I wouldn't want a company themselves doing their own quality assurance for this white hat hacking function, but hopefully their profitability would be able to fund such a function in an independent or not-for-profit framework. Thank you.

Dr. Bhatt: Dr. Shah.

Dr. Shah: Right. I want to specifically answer Troy's question about data drift hallucinations and also what to do with it, right? So I think this is like a no unknown, question that you're asking, right? If people who are submitting a new weather data drift was then they would've fixed it. So that's the big problem that everybody's going to face and it's going to be really hard to put that problem back on the submitter of the evidence because then they will have to do all this extra work to Dr. Miller's point

about not burdening the submitters with all this extra work. And neither do we want internally at the FDA people have legions of reviewers reviewing all this. It's just not pragmatic. So the problem is that if it should all go back to the label of the product, if you're labeling your product to do something very specific, let's say you're characterizing segmenting prostate cancer, Gleason grade 5 tumors for example, that's the label, that's what they're sending.

Then the data questions, there should just be a questionnaire to start with to say how many examples of each type do you think your model has seen for which your product is labeled? And if they say, we have seen 10,000 examples of this tumor, 5,000 examples of this tumor, 3,000 examples of this tumor, then at least we know the upper and lower bound of their data and their performance and we can at least anticipate where it'll go. Then also ask them the confession thing that Dr. Miller was saying, where do you think the model might drift in your evaluation of training it? And they might say, we have not seen too many Gleason grade 5 tumors before. So that's one thing. I think your life is significantly more harder when you have third-party foundation models because you really don't know what the underlying training data was on which the medical AI product was built.

And that's going to be a big problem to sort out. And my recommendation would be that we ask them that, does your underlying foundation model in your evaluation, has it ever seen the types of examples the product has been labeled for? If the answer is negative, then the significance of data drift is high. And if they say yes, then it's like okay, at least they have a notion. And then you ask them about the fine tuning of the foundation model on the specific product label and say how much fine tuning has been done and go from there. The second part was the hallucination part you asked, and that's a specific problem for text-based models as well, not just vision-based

models. So for vision transformers or their tokenization problems or for other models, we need to know what hallucination looks like for it to colon hallucination. That also goes back to asking a simple questionnaire versus all this evidence, where do you think this will drift and what do you think are the obvious stress points that we should be careful of?

Dr. Bhatt: Wonderful. Thank you so much. We are a little bit over time. So Mr. Tazbaz, is this adequate for now and are we ready for a break?

Mr. Tazbaz: I mean with that transition? No. Yes, it's adequate. Thank you.

Dr. Bhatt: Excellent. We'll take a quick 10-minute break. When we come back we'll hear from a couple more experts and the next question only has one part, so that should be good. Thank you all.

Sub-Topic: Risk Management

Stakeholder Perspective – Strategies and Controls to Mitigate Risks Associated with Gen AI Applications in Healthcare

Dr. Bhatt: Thank you all so much. That was a wonderful discussion by the Panel and we'll all slowly move to our seats to resume the meeting. I see people eating apples. I'm very proud of your healthfulness. I personally have a cookie, but it made for a good break. Thank you so much for taking your seats. If there are conversations that need to happen, please feel free to step out of the room for just a moment and then come back in if you are not actively participating.

Okay, first of all, I want to thank the Panel for their thoughtful recommendations in the previous session, there was a lot to go through and they really not only hit the highlights but went into great depth. We'll now resume the meeting and we will hear four more presentations after which the Panel may ask questions to each of the presenters. Our first presenter is Dr. Michael Schlosser, who is Senior Vice President of

Care Transformation and Innovation at HCA Healthcare. Dr. Schlosser, you may approach the podium. Oh, you're already there. Thank you.

Dr. Schlosser: Excellent. Thank you. And thank you for the opportunity to address the Panel. We at HCA Healthcare are excited for the possibilities that AI can bring to our healthcare teams and our patients. The opportunity to make healthcare more efficient and effective and a better experience is really a generational one. So we appreciate the FDA and the federal government's thoughtful approach to ensuring that common-sense regulation is advanced that supports responsible use of AI while not stifling innovation. I also want to clarify that I'll use some examples from HCA Healthcare in my presentation and while these are not medical devices, I believe the framework we'll use to evaluate them or we are using to evaluate them is broadly relevant. Just a little bit of context, which I think is important as I go into my presentation, HCA Healthcare is one of the largest provider systems in the United States. We have 188 hospitals across 20 states, around 2,500 other sites of care and we have the privilege of providing about 43 million patient encounters a year.

Why that context is relevant is when you start evaluating generative AI use cases in particular and you start thinking about things like hallucination rates and error rates and you start talking about 1% and 2%, the mass starts to add up really quickly. I'll give one example in my talk about nurse handoff, the handoff between one shift to the next in a hospital. We do that about 400,000 times a week, about 22 million times a year. So those 1% error rates start to look very different at that kind of scale.

So three real topics: two that I'll speak to in a little bit of depth and then one that I'll just make some quick comments on. The first is really around governance, which I think speaks directly to the topic at hand of risk mitigation. And I'll speak to the governance structure that we have stood up and are using and studying and learning

from across all of HCA Healthcare. And then the second topic is around a specific risk-mitigation strategy that we have used with some frequency, which is the human-in-the-loop and there's been a lot of conversation around that today and so I think a little bit of a deeper dive into that is really useful and warranted. Then lastly, I want to make just a quick comment on data and data use. All AI strategies are data strategies, and so I think that's an important part of the conversation.

So this is HCA's responsible AI framework: We developed this and put it into place about a year ago. We have a policy which is available on our public website that describes this in some detail. That policy went into effect earlier this year and really requires all of our 300,000 colleagues, including our practicing physicians and nurses and others, to engage in the responsible AI program if they're going to use AI use cases within the bounds of our healthcare system. So whether they're using it on their personal devices like we heard earlier today, people just pulling up Chat GTP on their phone to ask questions or whether it's an AI system that we have deployed or are planning to deploy, the policy really directs that the first step is to engage with the responsible AI program and go through this risk mitigation strategy before the AI is really used at scale.

The backbone of the strategy really is a risk-based approach. So as we've talked about a lot today, the types of AI vary dramatically from simple algorithms to these very large complex transformer based generative AI models. So a nuanced approach that really looks at risk and benefit and then understands whether we want to take on that risk and the benefit outweighs it. And then what are the mitigating strategies that we can use to decrease the risk is really the approach that we saw as the most successful one. I'm not going to go through everything that's in that wheel, that wheel describes our program, but I want to highlight a few of the key areas. You can see the seven

domains. I will talk about those on the next slide in a little more detail. Safe and secure, private, transparent, and again, I'll go into those just in a little bit more detail.

Those are the domains or areas of risk and it's our framework. There are many other similar frameworks out there that we use to identify what are the unique risks that any specific AI-application if deployed could create. Underneath those the individual wedges are sort of sub-domains so more detailed questions in each of those domains that help us really get to the specific nuanced risks that may be created based on the AI solution we're considering deploying. Around the outside of the model are various mitigating strategies or levers that we have to pull in order to try to mitigate that risk, so we talked earlier, the Panel talked earlier about re-skilling and education and how important that would be. And so you can see that's a key part of our strategy, machine learning operations, which really are the systems for what we would call here postmarket surveillance, but how do we continue to observe and manage models that are deployed at scale and ensure that they continue to perform at the level we want them to is another critical lever when we think about deploying these kinds of models.

And then the other areas are our ethics, our cultural norms. These are ways that we expect our team members and our colleagues to show up and engage in these processes and do things in an ethical way, in ways that are consistent with how HCA Healthcare practices in general. My last comment on this slide is that the team that represents this program that administers this program and it's a team of leaders and data scientists, really understand that their job is to enable the use of artificial intelligence, it's not to limit it. It would be very easy to manage the risk by just saying no, that's a pretty easy lever to pull, but that would again miss that generational opportunity we have to transform healthcare. So they understand their job is to identify the risk to

mitigate it and then only to say no if we feel like we're in a zone where we're not ready to take on a level of risk that a product potentially creates.

So this is the seven domains of risk that we study with each of our AI applications before we would pilot it or deploy it. Just to comment on that, many times the risk assessment really goes in multiple phases where we may only understand a piece of the picture of the risk it creates when we're just studying it on paper or studying other folks' use case or use of that product. You actually have to deploy it in a pilot and study it to really fully understand how these risks may show up, that gives you within a second layer of coming back to this process and then restudying and determining do we have the right mitigating strategies for a scale deployment versus a pilot deployment? So there really are no surprises here, in fact, I think this is very similar to the FAVES construct that the ONC has used though these don't organize themselves into a nice acronym, so the government couldn't make use of this framework.

But otherwise, I think very similar concepts. So safe and secure, we will establish careful and intentional guidelines to protect patient data against risk and ensure data is trusted and interpreted accurately. I'm not going to read you the entire slide, obviously you have it in the record, but each of these questions or statements describes to the business user, the end-user or the person responsible for this AI application, what it is that we expect of them as we think about how the application or the model is deployed and the mitigating strategies that we put into place. I would say that if you think about safety and security, privacy, really fair and impartial, these tend to be pretty technical. You're in the weeds there in terms of how does the model connect to our data systems and what are the ways in which those connections are governed and how do we ensure that the data is being transmitted in a safe way and that the model can't be used

to access the data, that it has access to very, very technical conversations. When you get into transparent and explainable and robust and reliable I think these are actually the more interesting domains of risk and it's where I'll spend the rest of my time discussing here today. These are areas where often we don't have the complete answer and we're still moving forward. I'm going to show you examples of how we're using AI, but I think the right answers are not completely determined yet in those areas.

Just one more point on this risk model that we have. So underneath those seven domains are actually 43 individual questions that make up a Risk Register. So those questions each map to one of those domains. Again, you have this, so I won't read it to you, but this is an example on privacy and so as we're considering deploying a model as a business unit or a user is considering using a model, they're asked to go through these 43 questions and determine how does their model map to this Risk Register. Is this a risk that that model particularly has? Is it a high likelihood that this risk shows up when the model's deployed? Is it low-likelihood? And then there's some data already in the register that suggests what the likely impact of this risk would be if it does show up. And then that all can be used, all 43 questions in conjunction to come up with a score so that we understand is this a low-risk use case, is it a medium-risk, is it a high-risk or a critical-risk use case?

And then we have a governance structure that oversees how we respond to those four levels of risks. So we have a very senior-level governance committee, which involves some of the senior leaders in the company like our chief lawyer, chief of ethics and compliance, and our chief clinical officer. The highest-risk use cases have to be reviewed all the way by those levels of folks before they would be used. If it's a medium or low-risk use case, there's faster paths to at least piloting and testing those use cases. But the rigor of actually going through this process and asking these 43

individual questions, understanding those seven domains of risk, mapping it to a level of risk, a bucket of high, medium, low, we believe is a framework with the governance structure to review that, that will help us deploy AI safely and effectively while not putting our patients or our care team members at risk in a way that we wouldn't want to.

A big piece of this strategy also comes very much down to use case selection. So a lot of what we end up wrestling with is do we want to take on this level of risk yet or not? And I'll tell you HCA Healthcare's approach to this is that we're very much tiptoeing into the use of generative AI and for a few reasons. One is I think generative AI, we still have an enormous amount to learn about these models and how they perform and why they do the things that they do. So putting a patient's treatment or diagnosis or other aspects of their care sort of directly at the end-result of a generative-AI model, I think is something at least our organization is not ready for. I think there's risk in taking on that too soon. Also in that if we take on those kinds of risks and we make mistakes and patients get hurt, we could end up setting ourselves back years before we ever get to the point of being comfortable doing it again. So this incremental acceptance of risk over time I think is a way that we can build trust with models, we can get people to use them in a way that's low-risk, but it builds comfort, it helps them understand how to use them, what they're good at, what they're not good at, which will make them better prepared to take the next step. And then the next step.

So this is just an example of what the risk register output of a generative AI model could look like. I'm not going to go through this all in detail here. I'm going to focus just on one bar or maybe two, which is the hallucination. So, of course generative AI models and in this case we were thinking about a use case where the model is producing an output like a medical note or a note summary. There's the risk of hallucination, which when you have the human-in-the-loop, we would rate that as a

moderate risk because the human is there to correct the errors. Then there is that risk of over-reliance on AI systems, which we have not really demonstrated yet, but certainly is something that we would be concerned about with these types of systems. So I want to focus just in my last few minutes here on this concept of human-in-the-loop, which I think is an important aspect of generative-AI models.

In fact, I'm going to make the argument that I think it's a required aspect of generative-AI models at this point in time. We often talk about human-in-the-loop as if it's just sort of one thing, which is the model produces an output and then we ask the human to review it, they make their edits and then they accept the output and then it goes into the medical record or whatever the next step of that model is. I think to do this really well is actually far more nuanced than that because at scale, again, if I go back to the 22 million handoffs, if the end users are not really engaging in this human-in-the-loop process in a robust way, you could end up with a huge number of errors making it through those humans and into the medical record. And then it's very hard to calculate what the downstream impact of that could be over time, especially as it accumulates over time.

So, here are six concepts that I think are important when designing systems that include human-in-the-loop as its safety measure. Promoting transparency in whatever way we can. We've had a good discussion about where you can be and can't be transparent in showing how a model made its decisions when you talk about generative AI, but there are lots of good transparency levers you can put into place, like asking the model to provide citations that show where it found the data that it displays. Displaying levels of uncertainty where it doesn't have a citation or can't show you exactly where that information comes from. Make sure that it's in the interface, that the user can see that level of uncertainty. Encourage active participation. So, designing the UI / UX in a

way that draws the end user into that activity. This is where your UI / UX designers are really worth their weight in gold. Provide training. We have a requirement in our responsible AI program that anyone who uses AI has to go through at least a minimum level of training. We have a series of micro-learnings they're exposed to which teaches them at least what generative AI is and what responsible use of AI looks like. So, they have some baseline understanding of what these models are and what they can do. Design for trust and calibration. And then last, implement feedback mechanisms. And I'm going to double click on "feedback mechanisms" here for just a second. I know I'm running short on time. I'm going to use this handoff example. So, we, partnering with Google, designed a nurse handoff tool to assist in that process. It's a very important time in care delivery, as all the clinicians know. We used a standard framework, the I-PASS framework. This is something that's been used for a long time. And we ventured to teach an AI model how to do this process, how to use an I-PASS framework to read the medical record and come up with a consolidated summary, a timeline of the progression of care and a prioritized list of tasks and key considerations for the oncoming shift.

We built into the user interface of that handoff tool the ability for the end users to provide direct feedback on five what we thought were the most critical variables in their assessment of the quality of that model: factuality, which speaks to hallucinations, coverage, organization, conciseness and helpfulness. And they would evaluate all five of these on this red, yellow, green scale. And then, there would also be a composite score where if they rated all five as green, then that would be a green note. It's important to note that these five features are actually, sometimes, in conflict with each other. So, you can have really good coverage, and you can have good organization if you allow the model a lot of creativity, but that usually leads to hallucinations. If you

make the model highly factual, it will often omit a lot of things, and its organization sometimes can be poor. So, balancing these things against each other is critically important to the usability of the model and the safety and helpfulness.

So, in this last slide, I'm just showing you one of the variables. This is factuality. We have this data for all five of them, but in the interest of brevity, I'm just showing you one example. What I think is interesting about this is if you look at model version zero, which was still at the time we did this work, which is this past summer, we used some of the best foundation models that were available. We tried multiple different ones, including Google, Gemini and OpenAI 4.0. And you can see that the model actually performed pretty poorly. And this was despite the fact that at the time the literature would suggest these models were really good at summarization tools, but when you asked it to do something just slightly different than summarize, and you applied a really critical eye to it, the nurse, and said, "No, I don't want just a summary. I want the critical elements, the most contextually relevant elements from the data to be presented," the model immediately did poor. And then, we had to go through a series of steps. And we didn't change the foundation model actually through this entire progression, we changed the system that sits around that foundation model. And I'll just editorialize. That system includes the person that's using the model; it's part of that system. And we created feedback loops where they would score on that red, yellow, green. That information was passed back to the system, sometimes directly to developers, to the prompt engineers, sometimes in the form of recording the output as green, the highest quality output, and using that as examples that we fed the model. And using all of those techniques in an iterative way, we got to the right-hand side of the slide where we were over 90% factual, and the nurses told us at that point the model was good enough. It was also over 90% helpful to be able to use in at least a production

pilot. This was a very controlled environment. And so, we've now moved this into the next step.

So, a couple of final thoughts and then I do want to mention my comment on data. So, first is-- I think when you evaluate these kinds of systems, it's very hard to detangle the model itself-- In this case, the model itself is really just a foundation model. It's Gemini 1.5 Pro. From the system that you put around the model, which includes the user interface, the feedback mechanisms, and in the end the people that are using the model. And so, my opinion is if we're going to evaluate the performance of a generative AI system, you have to look at the performance of the system, not just the technical pieces of the model.

I also believe that in order for AI to be successful, it actually has to be used. And so, there's a certain threshold that we may want to see before we were deployed something broadly. As you can see, we did that ourselves. But I think that we're going to have to come to an acceptance that the answers to these questions are never going to be 100%. And so, we're going to have to study how do you deploy a model that's 98% successful? How do you ensure that it's designed such that humans are monitoring that 2% in a successful way? And then, how are we returning that data that the humans are providing us back to the system and forming a continuous loop so that the model is constantly correcting errors and getting better? And that should protect against some forms of drift and other changes in the environment that the model may encounter. But just understanding the premarket performance and assuming that it will somehow provide safety in the real world I think is going to be incredibly challenging for generative AI.

My last comment on data is that these foundation models are incredibly powerful, but it's the context-relevant data that makes them useful in workflow. And so, these are

always, in my opinion, going to be a combination of technology companies who have the ability to generate these foundation models and folks like ourselves who have rich context-sensitive data. And it's the bringing those two together that's really going to create successful outcomes. And so, while I know FDA doesn't necessarily regulate or doesn't regulate patient data, I think making sure that we don't create dogmatic rules when it comes to how data is used that makes it impossible for us to do that ongoing context-sensitive training, I think it's also going to be important for success. So, I'll end there. Apologize for running over time and thank you.

Stakeholder Perspective - Narrow VS Generative AI: Risk Determination > Controls => Safe Innovation

Dr. Bhatt: Great. Thank you, Dr. Schlosser. Next on the agenda is Dr. Keith Dreyer, Chief Data Science Officer and Chief Imaging Information Officer at Mass General Brigham. Dr. Dreyer.

Dr. Dreyer: Thank you and thank you all for the opportunity to present. I'm a diagnostic radiologist, but I have spent the last 10 years of my career dedicated almost 100% to the creation and deployment of AI in the pre-Gen state and in the current GenAI state.

I'm going to talk about narrow AI versus generative AI with respect to risk determination, controls and safe innovation. I think it's important for this Committee to be able to get a good purview of what the FDA has done successfully and some of the challenges that have been in place through the narrow AI era, if you will, in thinking forward, as to some of these pluses and minuses might progress into this generative AI state. I want to look at technology, regulations, usage, and monitoring as kind of this life cycle. When the technology CNNs came out in the early 2010s, it was seen as a transformational technology that computers for the first time could do things that they

couldn't do before at a percentage that was acceptable. And so, that entered the foray of the concept of putting this into healthcare. I would say that this challenges the regulatory system in this country and several others around the notion of should we create a software as a medical device and treat these as devices? They're obviously not food or not drugs. So, if you have to have something that already exists, device probably makes sense. But I still think that this Committee should even question today if that rubric of a device is the right way to take a look at GenAI, like what was mentioned with Bakul Patel's recent article talking about this is more human-like. And so, I think some of the challenges that we're going to see historically looking back is that we maybe made this decision too prematurely. The adoption of AI in this narrow AI state was also somewhat limited, and I'll talk about that, for a variety of different reasons. And, as you probably all know, monitoring is almost exactly zero throughout the United States in this by device manufacturers as well as those that have deployed the technology.

So, I just want to talk about what happened through this stage of good technology that ran into some regulatory hurdles. First, I know the decision is: is it AI? Is it used in healthcare? If the answer is yes, it falls into HHS. If it's a medical device that the answer is yes by 21st Century Cures or whatever mechanism you want to use to be able to determine that, then it falls into FDA, this group here, and makes the decision falls into SaMD or Software in a Medical Device, and then gets weighted for risk. The reality is Software as a Medical Device is essentially all one risk's number two. So, there is no risk stratification, as Dr. Schlosser described what they're doing. And then, there's different ways to get them approved. Probably most of you are familiar with this, but almost all have come through the 510(k) pathway, being that these devices are predominantly imaging-based, where 75% are radiology, but upwards of high 80% are

imaging of any kind, optical, cardiac, etcetera. And they fall into these known categories. And these categories also kind of determine risk, but now at the level of specificity of risk, category I, II, III. That has resulted in 900 AI medical devices for imaging of any type, a thousand total, another hundred that are Software as a Medical Device for non-imaging. And I want to focus on these. This is really because that's what I do and also this is the predominant area that has been used for the last 10 years and approved and deployed, because I think it's a good lesson to look at it retrospectively to see some of the things that could happen going forward.

So, basically there are two different categories, despite the fact that there are multiple risks and it's really determined by-- These six categories are defined, but yet when it comes time to be able to impose a control to limit that risk, there's really only two choices. There's the concept of standalone performance testing, or comparative effectiveness or reader study. And that has created some challenges in the market today-- I'm going to put my glasses on. Challenges in the market today in that the cleared devices of those 900 look like this. About 13% are those high-risk, but high-value solutions. And the bulk of the products that are available today are these CADt, CADa/o and MIMPS-type solutions that had a lower bar to get through. So, it was much cheaper as opposed to millions of dollars to go through a CRO-approval process. It was thousands-- Hundreds of thousands and took months, instead of years. So, the result is there's a predominance of triage software and non-diagnostic solutions, and there's a very small number of diagnostic solutions available today. The result in that is-- And through the American College of Radiology survey of radiologists asking about adoption rates of AI over the last five years, it's about 2% today, and there's reasons for this. The FDA requires that triage not to show imaging results. So, because triage is seen as a lower solution and standalone performance test is okay, I believe the belief is,

or could be discussed, that if you do show images, now you're making a prediction and as opposed to just triaging the patient, but that also has an unintended consequence of making it much more complex for that human to interpret that. So, if I know there's a problem in the image, but I can't see where, and I haven't been told where, I'm naturally going to take longer and be more concerned that I'm going to miss it, but I won't know what it is that I'm missing. And since all AI products require human oversight with the exception of one-- So, that 900 require that. That triage AI makes the workforce burnout worse for radiologists. A very important point here is that if I see that there's-- if I don't see what it is that the AI is saying, but the AI is saying there's something there. It's going to slow down those that are doing the interpretation. Therefore, many experts won't use this type of AI, and that represents 87%. CMS has created fewer than five CPT codes for payment of AI. And there's no commercial coverage for FDA cleared AI products today. So, there's no incentive to deploy. That means that it's off the back of practice performance that needs to be purchased. And these same products in other countries are considered diagnostic solutions, providing improved patient care, but the AI coverage of diagnostic tasks in the United States is about 1%.

And how would I come to that number? If you look at what a radiologist does, a fully capable general radiologist, of which there are a few, because everyone's subspecialized, there's about 3,500 findings that you would have to find to be able to read all modalities, all organ systems. If you look at the landscape of AI in the United States today approved for CADe or CADx-- It's this limited number that are there inside of those boxes, as you could see predominantly in mammography, chest X-ray, etcetera. But there's about 42 of those 3,500 that are necessary, which is less than about 1%. Also, of note though, if you look at the chest X-ray as an example, seven findings

that are CADe or CADx for diagnosis are-- The same AI products in the UK and Asia provide full diagnostic coverage. Those CADe, CADt solutions here are full diagnostic solutions elsewhere in other sovereign nations.

So at the end of that, because all AI requires human oversight, but not by an expert, arguably if a non-qualified individual is using that AI to make an interpretation, it's arguably coming out sometimes with a worse diagnosis than if a human was able to do it completely independently, separate from the other lesser qualified individual that would be using AI to make the diagnosis. And we've had these problems in the organization-- At Mass General Brigham, in our organization. Limitations in SaMD risk classification in my mind are preventing AI from providing patient benefit at this point.

So, if that's the state today, what happens as this technological innovation transforms into generative AI technology? Clearly, again, an absolutely incredible technology that appeared on the scene just in the last few years. There is no regulatory pathway to have these medical devices approved. Dr. Schlosser approved or rather discussed different ways for non-medical device AI to be used, but there is a tremendous challenge right now to try and use GenAI inside of a regulatory or medical device, even if it's created inside of an organization. So, yeah, usage is somewhat high, and monitoring is still bleak. And what do I mean by that? Well, there's pathways for people to be able to go straight to the solution, like GenAI, drop an image in or do various things, that is happening today. And so, it's not this life cycle. It's like a direct line; a lifeline approach to this is what's happening today.

So, let's talk about this regulatory challenge now in light of narrow AI and also generative AI. If you look at the differences between the two, at least from my standpoint of how I see the challenges of regulation, which have clearly been discussed here, traditional neural networks are deterministic, they retrieve, they don't generate,

and they have narrow capabilities, whereas these are probabilistic, they're generative or can be generative, they can also retrieve and they have emergent behavior, unforeseen activities that they can perform. So, the question that I think is reasonable to ask is this notion of risk and that process that has been capable of controlling these deterministic, retrieval, narrow-capable devices with standalone performance tests and comparative effectiveness tests. Could that same kind of thing be applied to probabilistic generative behavior? I think the answer is yes. I think that you could just extend the risk curve and be able to use the same controls that were considered before with some caveats. If you look at this as all medical devices independent of the technology, but just looking at them as their intended use, their users and their risk, then I think you get a kind of good analogy. But the thing I would say, and I'll talk about this more, is I would not use comparative effectiveness. Running a CRO and watching the use of reader studies compared to standalone performance tests, I don't see the value. I don't understand the value in comparative effectiveness testing. I don't see that it does anything stronger than a standalone performance test, as compared to continuous monitoring and things you can do far beyond premarket testing. So, what I would do is look at standalone performance tests required for everything, as well as clinical validation. And I'll talk more about that in a second, what I mean by clinical validation, but then I would also look at the notion of ongoing monitoring. And those things should be based on and controlled by the amount of risk that the device imposes.

So again, if I look at this flow chart and look at these CAD solutions that are out there today, standalone performance tests handle these and comparative effectiveness tests handle these, but I would argue again that standalone performance tests could probably do adequate at that. These are all premarket testing with no monitoring. So, if you add GenAI onto this list and it falls through as a SaMD or SiMD, then probably, I

hope, the knee-jerk reaction is to just do comparative effectiveness testing because I think it's going to have the same challenge that it's had previously. And what I would do instead is start to impose standalone performance tests plus validation, plus the concept of monitoring. And now you have all these new devices that have to go through that rigor to various levels and degrees, but there's also-- As Dr. Schlosser mentioned, there are non-device GenAI applications that also need to be considered. I'll talk about those in a second, but those are not medical devices. So, somewhere up in the level of HHS, maybe not FDA, maybe that also needs to be managed and considered from a risk standpoint and for mechanisms of controls.

So, now we have devices that are not able to be used inside the United States, and we have devices that are able to be used inside the United States. And so, it's a very confusing time for providers like us to decide where the line is drawn, what these things are. And then how exactly do you enforce these internally and how do you deploy them?

Let me just use one example of voice-to-text transcription as well as diagnostic draft reporting. And it was interesting to see this used so many times as an example today, but let's just do that. So, text-- Just like we do speech recognition, as opposed to image recognition or image ability to create a text. So, the process is, as a radiologist-- And we have been doing this for decades. I look at an image, I dictate the interpretation into AI of some sort. Previously, they were RNNs, now they're more and more LLMs or foundation models. And it does voice-to-text and I get a report. That's a draft, I sign it off and I'm ready to go. What I'm talking now about is the notion of taking an image, putting that into essentially the same AI, but now I get a draft report out and now, as the radiologist, I look at the draft and I look at the image and I make an interpretation. And now that's how I get the report. So, transcription is considered to be not a device, but

transvision or whatever you want to call it, it just made up the word, is a medical device if you look at the rules, 21st Century Cures. But it's the same AI solution.

So, the question I start to ask is if transcription is okay to use and *transvision* is not okay right now to use unless it's-- As it's considered a medical device, and it has to get approved through a regulatory process to be determined, what about summarization? Today, there are applications that say, "You dictate half the report, and you let AI complete that report for you into the impressions". That has at least been deployed and used extensively throughout the country and has not had any kind of FDA restriction on it. So, I would say it's probably not a medical device, but unknown. What about quality checking? Others are talking about the ability to be able to say, "You dictate whatever you want and then I'll take a look at what you dictated. I'll take a look at the image, and I'll produce the difference between that to help to try and decide if there was anything that was missed." Again, unknown. So, these are things I think that this group has to decide as to what are the bounds and confines of generative AI that falls under the rubric of a medical device? And does that even matter anymore today? And then finally, consumer usage, where there's no claims and no regs, seems like it's okay to use. So, my biggest concern and what I see happening are-- Healthcare, yes. We welcome AI. Come on in, generative AI. Please, come in. And then, "Oh. It's a round peg." And all of the AI that's already there is going through a round hole, and this says, "Wait… I can use the consumer entrance." No claims, no regs, no problems. That's what's happening unfortunately. I know that no one wants this to happen, but this is what is happening until we find a better way to solve this problem.

I want to use diagnostic draft reporting as a GenAI case study of what I think was happening. This is how we're studying this up at our organization. So, if you think of this as the same model, it's doing both, then you can think of it like this. So, I,

basically as the radiologist, am looking at the image and I'm doing dictation, just like I have been for decades. That's being handled by AI, and it generates a draft report. What I'm talking about now-- Sorry. Then, I read that draft report, say, "Yes. It says what it's supposed to say," and I have a final report. What I'm talking now about is also adding the image into that AI so that it now can add additional information to that draft report so that I can do the same thing I've always done, but now that's a medical device. Now, if this was accurate, and this has been described in other presentations, it would add up to about 150% efficiency. We've seen no efficiency from the AI that's available today through the 900 products that are available to radiologists. This would clearly just improve efficiency. So, the question is how do we test this component for accuracy? By clinical validation and ongoing monitoring of this device.

So, Bernardo, Dr. Bizzo, talked about the healthcare arena, but basically this is a collaborative community in the form of the FDA's collaborative community concept to be able to create a solution that allows people to be able to look at these large language models that are for medical devices or not for medical devices and see if they're applicable to the environment that we have, to the data that we have, to the subject matter experts that we have, and the involving foundation models that are being used. So, what it does is it takes de-identified data and AI models and experts and makes them available here. In this example, it's a draft report. The important thing to note here is that, when I talk about a model and the instructions, the instructions are the prompt and the prompt is actually baked in here, right? So, it's not an open prompt. I can't ask whatever I want. This has been specifically trained and designed to be able to create a draft report. And so, I take de-identified data, I run it through that generator, that draft report generator, and I get an AI result. I get a draft report. As the expert, I look at that, and most important, I look at the input data to that model as well. So, now I see the text-

- I'm sorry, the image that the model saw, and I see the output that the model generated, and I rate that. I rate that from a range of acceptable all the way to unacceptable. And through that process of multiple experts with multiple data being able to run this, I can now start to extract discordance, performance, generalizability and bias, because we have ways to pull back to the data to say, "This subpopulation is biased," etcetera, etcetera. And so, the arena gives you this geographically and demographically diverse, embargoed data; unlimited clinical use cases (this one is the draft report generator); geographically and institutionally diverse experts, and expert evaluation of AI agents with input data. So, I can look at the input data and look at the agent's output data, and you wind up with these analytic evaluations of agents in real-world environments that help us to make a decision of whether we're going to deploy or not. Independent of FDA approval, we have to make that decision, and as well create performance ratings in a public forum.

This is what it looks like. So, in this particular example, I say, "Create a draft report." I can select the areas that I want to look at, emergency pulmonary, and I can now see there's clearly a fracture here. I can also see the output of the model that created that report that I'm now rating and evaluating. And I say, "I feel that this is at the level of-- At a Likert scale, resident three." I've got another example here. I read through the report. Another example. I rate it, etcetera. And now we have the ability to create these structured outputs where these stand and all of those statistics about these models that we need to decide to determine whether we want to use or are able to use.

The one thing that we've seen early in the deployment of this is the ability to reduce risk of these models. So, one thing I would petition the Committee to think about is if there were pathways to be able to solve this problem while you're still trying to solve the entire problem, but it allows models to be able to come out that are safe or low

risk, it'll help us all learn much more quickly how to go to the next ones that are higher and higher risk. I wouldn't try and solve everything at once. This is what I guess I'm trying to say. So, maybe if an open query isn't the best way to go or there's no clarity on open query, at least a closed query using LLMs might be something that's acceptable. So, in this case, the ability to create this information with the user and interpretation with an open query is pretty high risk. But what we've seen that the risk is much lower if you have fixed instruction prompts, if you show the input data, if the user is expert in that input data. So, if it's a radiologist looking at MRIs, if it's a pathologist looking at pathology images, as well as the AI. And the interpretation, the usage is monitored, then that risk drops considerably. So, fixed instructions, input data shown, user expert data and usage is monitored, is the big key. On this graph, it looks like this. You want to monitor that usage.

So, we see in this process of validation, of testing the model before it goes into deployment, it informs the manufacturer of potential defects that determines the site accuracy and generalizability; it lowers the risk of failure on site; it promotes early and safe innovation, and it provides a mechanism for ongoing monitoring as well. You can use the same solution for that.

I'll finally close with this thought. Is there a thoughtful way to regulate healthcare AI beyond a medical device definition? So, if you think of this complex drawing that's got all these different pathways is the way to maybe think of this to say performance tests, site validation and site monitoring are three controls that can be used for medical devices or should they really be things that are be able-- Accessible for devices, whether they're medical devices or not, just the fact that they do use AI?

Thank you very much.

Stakeholder Perspective – Safety from the Systems to Patient Levels: Risk Management for Large Language Models in Healthcare

Dr. Bhatt: Thank you so much, Keith. Our next presenter is Dr. Danielle Bitterman who is the Assistant Professor at Harvard Medical School. And then, we'll have one more presentation directly after that. Thanks, Dr. Bitterman.

Dr. Bitterman: Thank you so much to the Panel for inviting me to speak today. I'm Danielle Bitterman and I am at Mass General Brigham and Dana-Farber Cancer Institute, Harvard Medical School. And today I'm going to be speaking from my dual roles both as a practicing clinician, I'm a radiation oncologist, as well as a researcher in natural language processing. My lab is focusing on developing natural language processing, including large language model methods, to support and enhance safe clinical care delivery. And we're very interested in developing methods to improve the automated evaluation and robust evaluation of these models. Here are my disclosures.

So, a lot has been talked today already about how large language models, foundation models, are developed, but it is really difficult, as others have mentioned, to disentangle the quality of the foundation model from the safety and risks of a device. I really do think it's essential to understand and continue to monitor the underlying large language model that will be integrated into any device. So how do we get to the helpful and useful large language models that have really catalyzed this explosion of innovations and medical AI? So, today's advanced large language models are generally trained by taking large amounts of text curated from the Internet and having the model predict the next word in a sentence. So, most of this text is not medical; it's standard text. And from that we get a model that is able to transform and handle language. It can answer nonmedical questions pretty well.

There also is a lot of information about biomedicine on the Internet. And so, though language models happen to be trained on that in the process of their general training, and as we've already heard, this has allowed us to develop-- This has led to models that are large enough and turn on enough data that can now pass a variety of different medical benchmarks, including the US Medical Licensing Exam. So, it really is quite impressive, but there is still this outstanding question of how does performance on these benchmark evaluations in the general biomedical realm translate into safe and effective applications that truly help a clinician in practice and are safe for our patients? So, while large language models have a lot of potential, they also bring unique risks that require proactive controls. And the goal here is to balance innovation and safety so we can realize the benefits for patients and providers as quickly as possible, but also as safely as possible.

And again, the training processes through which we arrive at a large language model is the foundation for any medical device that uses that model's performance, behavior and risks. So, we already discussed language model pre-training. Most of these large language models also go through additional tuning processes, known as instruction tuning and our preference tuning. So, instruction tuning is when a model is given an additional fine-tuning dataset with instructions and the output of the desired instruction. They're tuned on that. That makes the model able to, of course, follow instructions. Preference tuning is a process by which a model is shown two different examples of a model output that have been human labeled for a preference. And while these make our models helpful, they also introduce new risks, such as leading to models that might be overly compliant, easy to jailbreak, as these models will often try and always give an answer, even if they may not "know the right answer" and that is risky for medical misinformation.

In our group, we're really interested in understanding whether we can start to understand the risks and behaviors of a large language model from its pre-training data. We carried out a study on a smaller publicly available large language model known as Pythia, for which the pre-training corpus is completely publicly available. And we were able to show that the pre-training corpora do reflect realities of current medical knowledge. Here you can see. These are the number of mentions of various diseases. I highlighted COVID-19 and infection in the pre-training corpora that these models are trained on. And you can see, as expected, COVID-19 and infections spike around 2019. So, that's encouraging. For this model, it is, at least at the time of 2019, up to date with general medical knowledge. However, we also showed that biases that were present in the statistics of the pre-training data led to biased diagnostic decision-making of the ultimate model that used those data to train on.

So, that's actually very valuable to know. If you're able to-- If you have access to a language model's pre-training data, you might be able to anticipate risks and its behavior in a more generalizable fashion than if you're relying on the inherently limited sets of benchmark datasets and red teaming prompting. However, this requires having transparency into the fine-tuning datasets-- The pre-training datasets and fine-tuning datasets used to develop a model. It also requires transparency into how the model is trained. And I do want to really emphasize, from a clinician and researcher's perspective, it is a lot of risk for the clinicians and patients to take on-- To onboard models for which we do not have enough information into how they were trained. I know this is likely going to be a challenge, especially from the larger companies where this information is currently proprietary, but when we're talking about clinicians' liability and most importantly patient's healthcare, it really is incumbent on us and we

have, I think, a really unique opportunity right now to require more transparency so that we can ensure that we develop these safe models.

Alright. Moving a little bit more into the translational side, there are risks at various levels of clinical application once we start to think about moving these models into clinic. These risks range from risks to systems privacy and security, risks arising from the device and also risk to the device, things like cybersecurity concerns; risks to patient safety and clinical effectiveness related to device performance, safety and impact and outcomes; new risks related to workflow integration for which we need better understanding of human factors, better ways to collect user feedback in real or near-real time and better methods for scalable monitoring; and finally, a lot of ethical and legal concerns regarding the transparency, equity, accountability and responsibility when one works with these models. And again, the behaviors and knowledge learned during model training and tuning modulate, and I would argue, really form the backbone for all of these risks.

In terms of systems-level risk and controls, some of these are unique to generative AI and some are more broadly applicable to software integrated into a healthcare system. But large language models might interface with many different points of patient data, clinician data within the Electronic Health Record system, as well as on the Internet and patients interacting with these models via their personal devices. So, data governance and clear definitions for data governance for the data input into the model and output from the model, as well as appropriate data masking, is important for understanding risk privacy. Security protocols, such as encryption, audit trails and cybersecurity protections, will be important from a systems-level. And, as we've heard a lot during today's talks, deployment controls for on-label use are really important. While pre-training and tuning form the backbone for a model's behavior, the prompting

methods are also very-- Are of course the kind of final step to determining what that model will output. So, role-based access controls might be one approach, if clearly defined, to help ensure on-label use, beyond simply role-based access controls, being clear about how a device is going to standardize, clean and ensure that the data input into the model, because those become the healthcare data which become the prompts, are consistent and will be important, and including issues about-- And ways to monitor for the type of data modalities, languages, and tasks that the model is being used for.

Now, underlying all of this is ensuring the adequate clinical evaluation of these models. Again, current benchmark datasets that are used to assess large language models, general biomedical knowledge, have clear gold standards and have reliable and automated evaluation methods, which is good. However, they do not reflect most of the really useful real-world applications that we care about and where there's a lot of excitement for advancing healthcare. So, I give one example of a real-world application, which is using large language models to help assist clinicians in responding to patient portal messages. This is a big source of clinician burnout, and many institutions are piloting this application. This is just one example of a question a patient might ask in their patient portal: "I've been experiencing hot flashes and night sweats for the past week. How likely is this a side effect of my prostate cancer treatment? What should I do?" Now, if I asked three different physicians or oncologists to respond to this message, you would get three very different answers. If I showed those three oncologists a draft response to that message, you would likely get three very different assessments of the quality of the response. And this really makes it hard to assess language model performance. Currently, there are no or very few gold standards or reference sets to automate the evaluation of such generative-type outputs and no way to reliably automate these evaluations. Human evaluation still really is the backbone, but

as I just mentioned, it has its own complications and requires clear definitions and transparency into how those evaluators were instructed to evaluate the quality of a model's output.

Okay. So, because of the limitations of existing evaluation benchmarks and the practical challenges of evaluating a small set of exemplars, a small set of inputs for your given task, I really do think that a multi-level holistic evaluation is needed to start to understand a language model's risk profile. So, one initial approach is to start-- Maybe understanding the general safety of that model. In the general domain, there are benchmarks to evaluate how likely a model is to be truthful. How likely is it to provide-- To produce toxic language? Evaluating these general harmful behaviors might start to give us a baseline understanding of the more explicit riskiness of a model that's being integrated into a medical device, but it's not adequate on its own of course. Our lab was interested in understanding just how these benchmarks might be-- The extent to which they might truly be reflecting a model's "knowledge" versus just kind of stochasticity or whether the model just was pre-trained on data. So, we did just about the simplest thing we could do. We took the USMLE exam, the benchmark most commonly used to evaluate a large language model's clinical reasoning capacity, and we switched all mentions of a brand name to a generic name of the same drug and vice versa. Now, when we asked-- So, this was an example. A mechanism of action of leuprolide is GnRH agonism. That's the correct answer. When we switch those two terms referring to the same drug, the models often responded differently. And that's despite the fact that those models in most cases were able to accurately match the brand and generic drug as being the same. So, benchmarks themselves need to be audited. This might be because the models were trained and overfit at this point to the USMLE Medical Licensing Exam. But it also speaks to the fact that language models "reasoning" or "knowledge",

and I'm sorry I'm anthropomorphizing a little bit, is very different to how humans learn and reason. Language models are probably much better than humans at memorizing huge amounts of data, but they have flaws in knowledge manipulation that humans may not have. And this complicates using benchmark dataset or testing credential-- Existing credentialing methods to monitor language models' knowledge because you might really just be testing memorization and not a true understanding of the concept.

We, in conjunction with several other NLP researchers internationally, have developed the TRIPOD-LLM reporting checklist to provide standards for what should be reported in biomedical research according to large language models. We lay out information that should be reported regarding the data used to develop the model or whether that researcher did not have access to information, to that data, how the model was developed, how the evaluation was carried out, who carried out the evaluation, and what were the guidelines by which the evaluation was conducted, and finally what were the results and endpoints for that study? Now, speaking to the kind of open-ended nature of large language models, we ended up using a modular format that adapts to different use cases, but this is one approach that we've arrived at to try and develop a more consistent method to review-- To evaluate these models.

Now, finally, as we've seen, many AI models work incredibly differently *in silico* than they do when integrated into a busy, complicated, messy clinic. When large language models started to be rolled out into EHRs for patient portal messaging, we were really interested in understanding how that may impact clinician reasoning and risks when used with a human-in-the-loop. So, we carried out a preclinical study where we asked oncologists to respond to patient messages asking about new symptoms related to cancer, things that we often respond to in a patient portal. First, we had the oncologist's manually written responses as usual. Then, we had GPT4 draft response,

and asked the oncologist to revise it until they thought it was appropriate and safe to send to a human. And what did we see? We saw that, in most cases, a language model response was safe and in about 60% of cases, oncologists thought that the draft could be sent to a patient without additional editing. However, there was a small, but clinically significant risk of severe harm or death without editing. So, that's concerning. However, if humans are affected at overseeing these models, we could reduce that risk and potentially it might be a safer implementation method. That said, we did see evidence of automation bias and over-reliance that could expose patients to those rare or more harmful risks of a model. So, we saw that clinicians-- The content of responses was significantly different when clinicians manually wrote the response, compared to when they use the LLM draft. The LLM just tends to reflect quite-- Tended to closely reflect the LLM draft, suggesting that physicians might be relying on the LLM's clinical reasoning more than just using it as an efficiency assist. So, even these more administrative approaches might impact clinical reasoning in unexpected ways.

So, when we're thinking about workflow integration controls, it will be really important to understand and to see evidence of how medical devices that use large language models impact effectiveness and safety when used as a part of the human-machine team. Automation bias and over-reliance are common and are likely to increase. And we can't rely only on human oversight to control these methods, to control these risks.

There are various different approaches to post-deployment and real-world performance monitoring ranging from regular quality assurance, prompt versioning and controls, ongoing audits, monitoring for off-target use and evaluation for shifts and how a user interacts with a model. Are they spending more time editing a response? Are more words being added or deleted from a report that's generated by an LLM? Those

may give us early hints of faulty performance and devices that integrate these controls are likely to be safer.

I'm running out of time so I'll skip over this and just touch on some of the ethics and legal risks, which is really important for clinicians and patients to feel they can trust a model. LLMs come with a host of equity concerns. They learn biases from the model they're trained on. Humans and machines may introduce new biases together, and they perform less well on lower resource languages. Transparency will be essential for ensuring trust from both the patient and clinician standpoint. Patients and clinicians really do deserve to know when they are interacting with a large language model that might not be always clear because some of their output is so human-like, and when their data is being used to further refine a model. And finally, clearly defining who was accountable and responsible for any risks and errors of a model is an important risk control. If a clinician doesn't know that they're going to be accountable for the output, they may not use it in as safe a way as is necessary.

We've defined some levels of human-in-the-loop previously, drawing from the Department of Transportation, categorizing models from assistive all the way to fully autonomous, which might be a helpful guide for further defining this.

So, in conclusion, language models have potential to advance health but have risks at multiple levels of healthcare. Challenges persist in scoping, monitoring and mitigating risks. And human-computer interaction modulates the benefits of risks and need to be investigated and risk-mitigated. There are several emerging approaches that should be considered when refining and developing oversight methods, including automated risk and performance assessments, interpretability methods and innovations in usability design. And the goal of this overall more measured approach will be to facilitate a durable and sustainable implementation of AI that advances human health in

the long-term in a way that patients and clinicians feel trust and comfort with our changing healthcare system. Thank you very much.

Dr. Bhatt: Thank you, Dr. Bitterman. We are going to forego our next break. Everybody gets 30 seconds to stand up for a second and not get sciatica. Do not leave your seat. Please just stand up and stretch. And while you're doing that, Gabriella Waters please come on up. Artificial Intelligence Evaluation Research Associate at NIST. You will have 15 minutes for your presentation.

Stakeholder Perspective - Risk Management for Generative AI-Enabled Medical Devices

Dr. Bhatt: Alright. Sciatica now being prevented. Everybody, come on and have a seat. Yeah. Okay. We now welcome Gabriella Waters, AI Evaluation Research Associate at NIST and the Director of the Cognitive and Neurodiversity AI Lab. Dr. Waters, you may begin, and you may notice me using two arms when things turn red up at the podium. So, my apologies if I start waving at you.

Dr. Waters: I will just wave back. Hi. My name is Gabriella Waters. Thank you so much for this opportunity to speak. I was furiously deleting notes as the presenters before me were going, so this will be quicker because we were kind of walking in the same lane on what we wanted to talk about.

I don't have to tell you what generative AI is, but as a theoretical AI researcher, as someone who develops these models, what I want to say about them in response to some of the comments that were given earlier is that LLMs are a kind of technology that you should be more amazed when they get it right versus when they get it wrong. The probabilistic stochastic nature of them lends them to furiously confabulating. So, the example given by Dr. Bitterman where the generic drug had the reversal of the names is very logical when you consider that it is rolling a die on what the next word should be.

It is only used to seeing the words in one order. So, when it's not in that order, it's rolling a die on what the next word should be. And so, when you have a model that behaves in this way, it's very challenging to deploy it in a high-risk situation, like in a clinical setting. It's also sort of risky to do any of this testing post-appointment; that is too late. All of the testing needs to happen beforehand. So, I want to put a good order of operations in place before we continue on through.

So again, I'm a theoretical researcher, so I'm going to always champion innovation and how we can move the technology forward and what that's going to look like. So, generative AI is really good at helping us to do personalized treatment planning, helping to enhance that medical imaging capabilities. We've seen a lot of this work in diabetic retinal neuropathy in terms of how models can actually perceive images and make diagnoses and look at decision support. We've got lots of examples of that. I'm not going to go through it. I told you I promised I wasn't going to go over the same ground, but there are definitely risks that we need to consider: data quality and bias, harmful bias. As a developer, we like bias in AI. It's the harmful bias that we want to get rid of. Good example: if I am looking at a dataset and I only want women in the dataset, the men in the audience might feel that's really biased. But if I tell you it's for a women's clinic, then that model needs to be biased. So, bias can enhance performance. It's when the bias leads to harmful outputs that we have to be on alert. So, just making sure we understand how the model is biased and under what situations that bias emerges, and then looking at the risks associated with that. We want to look at model transparency and explainability, and I keep those two separate. People who hold IP do not like transparency. So, we want to do a tradeoff between transparency and explainability. Full transparency means I know the entire suite of your ensemble. I know every single algorithm. I know how its architecture looks. Real-world

transparency means I know when a model is being used. Explainability means that any person in this room can tell you how the model arrived at its decision. That's important for building trust; it's important for deploying things in the real world. And we're also looking at performance variability in the real world. LLMs and generative AI are non-deterministic. So, you heard from Dr. Dreyer about deterministic narrow networks. We know that if we pull a piece of information from that network, it's going to look the same every single time. In generative AI, that's not the case. You can ask the exact same question on 40 different machines and get a slightly different answer every time. And if that tiny bit of variance leads to a misdiagnosis even 2% of the time, that's dangerous. And so real-world performance has to be something that's evaluated and tested for. At NIST, we have the AI Risk Management Framework. So, Dr. Dreyer also mentioned validation. We also need reliability. It's got to be reliable, and it's challenging to do that if you haven't tested for it beforehand. It's difficult to go back once it's out in the wild and then try to redress some of the challenges that you're seeing within the model.

Dr. Bitterman mentioned the governance controls. We definitely need some form of compliance as these models are developed and deployed and implemented and what that looks like. We need some sort of Review Board. As a researcher, I have to submit to the IRB; as an AI developer, I do what I want. That's not right in terms of clinical applications of this technology. So, these are things that we need to consider.

We talked about training, validation and controls. Robust training protocols: what does that mean? It is literally a free-for-all. You can develop your model in any way that you like. Do you remember there used to be a slogan "have it your way"? That's literally what it is when we're developing models. We just do whatever it is and then present it. We need protocols for this. We need those rigorous validation

procedures. We need to have TEVV as a standard part of the AI implementation process. The design process has to be there. Testing, evaluation, validation and verification.

Feedback mechanisms. It can't just be-- I have the thumbs up there for a reason. It can't just be helpful and harmful. Binaries do not provide context. Why is it helpful? Why is it harmful? What happens when it's good? What makes it good? The model can't learn outside of the context, and it can't update with just a thumbs up or a thumbs down. We love Likert scales. We really do. What else can we use to leverage training these models for more robust performance? We need user feedback integration. How do we get all of the developers who are developing models for people to actually like people? Because at the end of the day, people are interfacing with these agents. And the number one complaint I get when I bring this up is, "People are messy." Like, "Yes. That's why you need to test with the actual people." This shouldn't be a radical idea. We need to get that continuous feedback integration going so that we can understand the surfaces for which a risk can emerge, what the impact is of that kind of risk, and then what's the mitigation strategy for it? If we are only observing these things after something has gone wrong, we are again too late.

Postmarket surveillance and performance benchmarking. This is where I usually take a step back from the microphone and take a deep inhale. Benchmarking does not tell us about risk. Benchmarking tells us about capabilities and performance, and those things are not concurrent with risk. That just tells me that the model is doing what you said it was going to do. And in the process of doing that, it's making misdiagnoses. It's making judgements that are inappropriate. It's culturally irrelevant. It's doing all these things, but its F1 score is great. Those things don't tell us enough about the context of the use of the model. And so, we want to make sure that we're not just hung up on

KPIs, that we're also looking at how robustly we can measure the risks and not just the performance.

So, I'm trying to go fast. I'm going fast. So, trends. I am going to fly through these slides because you don't really need all of the details, but you need to know that people are thinking about how they can integrate generative AI into everything. As an AI researcher, I get two questions. What is AI? Is it chat GPT or is it the Terminator? Those are the two. So, everyone thinks that generative AI is AI, and businesses are trying to capitalize on integrating these systems into whatever use case they have. And so, you're going to see through these slides that a lot of businesses are looking at an adoption curve to really drive home how they can get the best ROI on GenAI. It's not great right now, but the promise of a robust return is there. And so, we're seeing a lot of interest in these areas. So, I'm just going as fast as I can. Looking at how we can scale up, that's another thing.

Theoretical researcher hat is on. What is on the horizon? Lots of potential use cases for what the patient will experience in terms of how GenAI will touch their healthcare continuum as they journey through the system. From pharmaceuticals to their providers, the papers, the medtech, the services, public health agencies, all of these areas are touched by GenAI. We're going to start seeing mobile devices as medical devices. We're already seeing apps that are already on board. Most of these devices that can track things like your heart rate, like your respiration, all kinds of-- You have things that will alert you if any change happens from your baseline. So, we're already starting to see this. And we're seeing a lot of online doctor interactions, online medical professional interactions. And the risks are-- Again, we heard about over-reliance, but we're also thinking about that as a two-way street because the more distance we place between the healthcare provider and the patient, the more the patient may distrust the

healthcare provider's advice, the more they may have trouble adhering. So, we have to look at how these systems can augment, instead of supplant. So, there are ways where you can have a LLM that is parsing and summarizing information that's being pulled in from the online portal, or if it's in a mental health situation and someone's submitted a report overnight and all of these kinds of information. But you still need to do that as part of a team. So, it's you get this information, it's digestible, and then you're able to follow up as a human. So, we need the human-in-the-loop. We need the human feedback. We need to very critically test and evaluate these systems. We need to understand their mechanics. If I rolled a die in front of you each time I said a word, you would not trust that what I was going to say to you was going to be coherent. But that is the nature of generative AI. It's a dice roll, every single word. And as long as we have these systems in this way, and as long as the nature of them is so opaque, it's really challenging. And I'm not saying Black Box because I love Black Box models. I love them. They're really good at finding patterns that humans don't understand. Opaque models are not really good at helping us understand why something is going wrong. So, it comes back to the explainability, it comes back to the testing so that we can understand. I have successfully finished early.

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

Dr. Bhatt: That was absolutely fantastic. All right. We are going to take a couple of minutes to ask some clarifying questions of the speakers that we just had. I'm going to take a minute, even though we're over time, to just remind us of where we've come from. So, the things we've heard are the following. We must create a safe infrastructure to employ the equitable use of generative AI devices in healthcare with guardrails and a robust evaluation mechanism. We are not trying to inhibit but to promote safety. There has to be strategies and controls to mitigate risk associated with GenAI applications,

governance frameworks, designing model systems for safe deployment and evaluation, and mitigating strategies that range all the way from ethics to reusability to cultural norms and re-skilling. We then learned that evaluation of the performance of an AI model involves evaluating the performance of the system and all its parts, including its deployment, its users, its interpretation, its iteration. And critical variables within a framework should include factuality, coverage or admission, organization, conciseness and helpfulness. But we very recently learned that benchmarking tells us about capabilities and performance, not necessarily risk. So, for risk mitigation, we do want to think about user feedback for GenAI model improvement and educate users about AI limitations and potential errors. Finally, we need to train clinicians and patients, when applicable, to use these devices because human-in-the-loop is essential, since at this stage we're looking largely at qualitative outputs, and we have few automated measures that assess closeness to truth in the absence of a human expertise. Equally important, however, is role-based access to AI models. If this AI model is not in your division, in your area, should you have access to that? Automation bias and over-reliance complicate the good part of human oversight. However, therefore, human training is still, again, essential for the success of the system. Lastly, we learned that premarket testing to date includes standalone performance testing or comparative effectiveness. However, previously, it was deterministic, involved retrieval and narrow capabilities. GenAI, we've heard quite a bit now, is probabilistic, generative and it has emergent behavior. So, maybe we look at standalone performance, clinical validation, and ongoing monitoring particularly. Lastly, we learned that generative AI, I love this, is furiously confabulating, a die roll for the next best word. So, we need to focus on real-world transparency, explainability. And real-world performance variability must be

evaluated prior to deployment in the premarket stage and not be relied on only in postmarket surveillance.

So, I think that is what we heard today. With that reminder in mind, do we have questions for the panelists? I will ask our team here to please keep your question limited to only a question because after this we have a chance to discuss question number 2 where you can give your opinion on the matter. So, specific questions. We'll take five minutes for questions of the Panel that just spoke. Dr. Soni, who are you asking your question to so that they can come up?

Dr. Soni: The question is for Dr. Dreyer, but really thought-provoking commentary from everyone. Dr. Dreyer, you really made us recognize some of the complexities of a lot of different pathways. You highlighted something that was really important, which is one of the outcomes that potentially should be evaluated is efficiency. I'm curious if you can expand on that a little bit more, and how do you measure it? How do you compare it to non-users of GenAI? And how do you continue to monitor it on an ongoing basis?

Dr. Dreyer: Yeah. So, this is a great topic and it's going to take longer than a couple minutes. So, I'll just paraphrase. I think I question if it's the FDA's job to measure efficiency because I think the market should do that. With a thousand devices today, people look at these solutions and say, "Is this making me faster? Can I afford to purchase this? Is there some return on investment here?" And that's usually through efficiency. It's assumed that FDA measures the safety and efficacy of this, makes sure that this is a safe device. But I think where reader studies get challenging is to try and demonstrate that someone is more proficient or efficient in the process of with versus without. I just don't-- I see that as something that we have to do every time we choose a

device. So, I don't know-- It seems very redundant and very expensive to try and do that *a priori*, as opposed to at a particular facility.

Dr. Bhatt: Dr. Dreyer, will you stay there? Do we have other questions for Dr. Dreyer specifically? Yes.

Dr. Radman: So, you talked a lot about your, I guess, advocacy of postmarket monitoring.

Dr. Dreyer: Yeah.

Dr. Radman: And I think we've talked a lot today about model drifting. I think we all agree that they're going to drift. So, it has me thinking about that fact and the FDA's framework of, as you mentioned, class II devices being the most common, which use substantial equivalents.

Dr. Dreyer: Right.

Dr. Radman: So, if you take a new device that has validation data, like Timeway this year.

Dr. Dreyer: Yeah.

Dr. Radman: And then you're comparing it to a previous device, the ecosystem which is tested is different. And that's the whole-- It's one of the components causing data drift.

Dr. Dreyer: Yeah.

Dr. Radman: Right?

Dr. Dreyer: Yeah.

Dr. Radman: Model drift. So, do you-- Considering your position on the importance of data drift and continuing monitoring, do you also agree with the substantial equivalence framework that it's equivalent to compare to a device at a previous time? Or should you do a head-to-head bake-off at the same time?

Dr. Dreyer: No. That's a great question. I feel like everyone goes through-- Everyone takes the easiest way to market. Right? And so, an equivalency test is what everyone tries to achieve. I've looked at some equivalency tests and looked at the two devices that are considered equivalent. I don't consider them equivalent, but I'm not in a position to make that decision. So, I think even beyond what you're saying, there's other reasons why I don't know that that really works well. And I think in GenAI time, it's going to be much more complicated to try and figure out that equivalence. The thing that I would say is, in my space-- And I try to remind folks that these devices are not acting on humans; they're acting on data that's coming out of a device that came from a human. So, for example, the 900 devices that are being used today in the old-fashioned narrow AI don't drift. These are deterministic, but the data that they're looking at are coming off of CT scanners, MRI scanners, EKG machines. Those things are changing all the time. So, the data is changing all the time. We have 2,000 scanners at our facility and none of them are the same. So, the concept of trying to have one device work on every single scanner, another device, that then acts on humans is just not realistic. And that's why I think that validation has to take place, as opposed to comparative effectiveness or premarket of anything, because it's a moment in time with fixed experts and fixed data. And I just don't think that scales, especially in the generative AI space.

Dr. Bhatt: Great. Thank you so much. Ms. Miller, who is your question for? And a reminder that we are focused on risk mitigation and management with our current panelists.

Ms. Miller: My question is for Dr. Schlosser. I think I said the name right. This is a very quick question. What is the intended use for your model? I'm not sure you described that well or I missed that or-- Exactly, how that's going to be used, how it's being used.

Dr. Schlosser: You're talking about the nurse handoff model that we talked about. Well, so, let me first clarify that I do not believe that model is a medical device.

Ms. Miller: Okay.

Dr. Schlosser: And so, we've not specifically defined an intended use the way that the FDA would, but if I had to come up with an intended use, it would be to facilitate the handoff of a patient from one shift to the next.

Dr. Bhatt: Great. Dr. Kukafka, who is your question for?

Dr. Kukafka: I think it applies to all three speakers, but whoever wants to answer it.

Dr. Bhatt: Since you're there.

Dr. Schlosser: Okay.

Dr. Kukafka: Okay. So, I think all three speakers talked about the need to train users and patients to mitigate risk. And as someone who's worked in more generally biomedical informatics and more recently AI, training clinicians is next to impossible because they're busy. I mean, we just can't do it. And certainly, training patients, all patients, seems less than feasible. So, I just wanted some response to that. And what's the strategy to do this training to mitigate risk? How is that actually going to happen?

Dr. Schlosser: Well, so, I would say that those two populations I would view very differently because I agree. I think if we are thinking about direct-to-patient AI models that don't have a clinician-in-the-loop in between the AI and the patient, then you are going to have to deal with the patients, which is a very large diverse population with different understandings of what that model is. And I think that's one of the challenges that we have right now about thinking about direct-to-patient generative AI solutions. It's that we can't expect to train them and educate them on how to use those models, which is one of the reasons I think that's high risk at this point. On the provider side, I see this maybe a different way, which is I think providers who leverage AI and

understand how to leverage AI are going to replace those that don't. And so, I see this as an opportunity more than a challenge. I think that it should be a barrier. In our system it is. You can't use AI applications, at least in the ways that we can control the use of AI, if you haven't completed at least minimum training. And I think those who choose not to go through that training and don't use the AI will over time become obsolete.

Dr. Bhatt: Great. Thank you. Dr. Elkin. Last comment and then we'll move to discussion. Our last question rather, pointed question.

Dr. Elkin: Yeah, so question. Two questions. Sorry, I apologize. Two. One is several of you mentioned that you had a risk assessment survey that you use. Can you share them? And then the second question is, and something I've been wrestling with-- People here were talking about different models for different modalities and that makes sense of course, but there's also the issue that these models may not work equally well for all populations and I wonder if we can find a highly accurate model that works equally well for all populations or we need to train many models and use the right model on the right population person. I don't know if any of you can respond to that one.

Dr. Schlosser: I'll grab the first one. I'm going to let someone smarter than me do the second one. We have shared our risk assessment survey and strategy and in fact I didn't mention in my talk, but we're actually collaborating with the FDA on studying that risk framework to determine if it actually achieves the outcomes that we think it will achieve. We think it's a good framework. It's been used in other industries, but we need to really study it. And so, we're going to take a number of AI use cases through the framework, study what the impact was on our system and our outcomes, and then, hopefully collaboratively with the FDA, publish those results. But we're happy to share it in its current form as well.

Dr. Dreyer: I'll go next unless someone's smarter and answers for me as well. This is a fundamental question I just think that you've raised here. It's that when we create AI, sometimes we create it as a product for the United States and sometimes we create it for ourselves. It's much easier to create a solution for a smaller population and you can get much higher accuracy by doing that. So, the question to this group is: is the concept of fine-tuning at the site level something that should be considered? Because if you think-- The goal is to try and have high accuracy devices. You're probably going to get them higher accuracy if you're focusing that onto a population that you're going to be using it for.

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

Dr. Bhatt: Okay. It is now time for Panel deliberations and the discussion of the FDA question number 2 about risk management and mitigation. Ms. Shick, if you'll please present the next question. Fortunately, it only has one part. We will have 10 minutes for focused discussion amongst the Digital Health Advisory Committee. We will then have 10 minutes for the FDA to ask us clarifying questions. I left a little wiggle room that can extend to about 15 minutes. Ms. Shick.

Ms. Shick: So, the question here is: "Risk management: What new opportunities, such as new intended uses or new applications in existing uses, have been enabled by generative AI for medical devices, and what new controls may be needed to mitigate risks associated with the generative AI technologies that enable those opportunities?" For example, controls related to governance, training, feedback mechanisms, and real-world performance evaluation.

Dr. Bhatt: Great. As you are putting up your placards, I will start with comments that were made earlier that probably belong in this section. To begin, the Panel generally believes that the output of generative AI-enabled devices should be

categorized according to risk, as we heard in the earlier discussion. Whereas this may be challenging, the basics of direct acute care provision, clinical diagnosis or triage of risk level may allow for differential stress testing suggestions by the FDA. I thought that comment belonged here. A centralized database was mentioned earlier today. It would be essential, in generative AI, to build in a method for users to report errors for risk management, similar to the Adverse Event database with particular attention to bias detection and hallucinations. Earlier, we also mentioned, despite FDA labeling and marketing requirements for device companies, even the best guardrails may not prevent natural human drift when it comes to generative AI use. Therefore, it's essential for us to think about postmarketing surveillance when we talk about off-label use reporting for risk mitigation. And lastly, human-in-the-loop feedback was essential. It allows promotion of transparency using explainable AI to understand AI decisions and to demonstrate confidence levels to inform users of the reliability of outputs. Those were the four comments I pulled from earlier from our team that I thought perhaps belong in risk management. Dr. Elkin.

Dr. Elkin: Thank you. So, I think that first we have to realize that there's at least two kinds of risks. One are errors of omission and the other are errors of commission that happen with these models. I think-- In terms of the new intended uses, I think the biggest one that I think most of us have talked about is that these present a new way of presenting information that we accumulate within a model that seems more humanistic. And because it seems more humanistic, it gives the impression to the user of real intelligence, whether there is real intelligence or not. I think there's lots of work being done toward making these more intelligent systems, but the truth of the matter is that the impression to the user that these are intelligent-based on their conversational capabilities makes them seem more authoritative in the way that they deliver

information. So, even if that information is available from other sources, the impact in on-prem and toured the information may be actually appreciated differently. So, I think that governance needs to be done at many levels. Obviously, local governance is probably the most important thing, rather than national governance of these things because value sets are often local. And I think we heard that by some of the speakers this afternoon. I agree with Dr. Kukafka that, although we'd love to be able to train people to use these, doctors are not-- They're very busy and it's not likely to be able to enhance things because in-- We know from EMR implementations that if you plan on your EMR requiring physician training, that it just won't work, and the best approach has always been at-the-elbow training. So, having clinician champions and clinical informaticians around that people can go to when they're not sure about things may be quite helpful.

Feedback mechanisms are expensive, and you don't know who to take feedback from in terms of the model. So, there's two ways to think about that. How you improve the model for the feedback, but then of course you just want to mirror best practice. Long time ago, when I was at Mayo, one of the top radiologists said, "Peter, what I want you to do is come up with a model that would tell junior clinicians exactly what the five smartest people in our department would've done with this case and that would help them to learn more rapidly." And I've grown to appreciate his comments over the years.

And then, how we use this as an educational tool to increase the quality of our care across the board. So, we often talk about best practice, but we still have a problem with the bottom. Right? We want to raise up the bottom to minimal standards of care, but then we don't want to stop there. Clinical guidelines do that. We want to get beyond that and move closer and closer asymptotically to best practice. So, I would look

forward to people's thoughts about how we can accomplish that with these new technologies.

Dr. Bhatt: Dr. Clarkson, I'll have you go next. I just want to make a differentiation again. There is generative AI that is inside devices, which is what is our purview here. However, appropriately we have discussed after Keith Dreyer's presentation that there's a lot of other generative AI that is affecting clinicians and patients. I understand that. But if you could maybe start with the GenAI device part and then if you have thoughts on the other, we'll record that as well. That would be helpful for anybody who goes next. Dr. Clarkson.

Dr. Clarkson: Hi. I'm Melissa Clarkson, the Consumer Representative. I want to pause a moment and ask: risk of what? So, our first presenter I believe had seven categories and one of them was data security and such. And those are all absolute valid risks to pay attention to. Risk of patient harm I hope is up there. As a patient advocate, we have a tense relationship with the word "risk management" because we know how this plays out in the real world, which is first try not to harm patients, but if you do, it becomes risk management of financial liability and institution reputation. So, with that said, I see an increased risk for lack of accountability in healthcare. It's already really difficult for harmed patients to get accountability, but if there's now generative AI, which seems very trustworthy and authoritative and your hospital has implemented it and it came from an expensive company and you're a hurried doctor and you used it and now a patient is harmed, tracing that accountability and getting accountability for harmed patients, I see this as an increasing problem.

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

Dr. Soni: So, in thinking about the feedback mechanisms in real-world performance evaluations and controls over that, perhaps one of the phrases that has been

unspoken, but on many people's mind, has been the learning health systems as enabling a lot of this generative AI when implemented as a medical device. How is that going to continue to inform evidence-based practice and practice-based evidence as more intended uses develop? And I think that may be an opportunity where-- Again, different agencies within HHS have championed learning health systems within FDA for ongoing monitoring. Where does that come in? An ongoing monitoring of real-world performance evaluation is easier said than done. But if every implementation is going to be better informed by local validation, then perhaps that ongoing local validation, and some sort of an audit report may be helpful.

Dr. Bhatt: Dr. Miller. Ms. Miller.

Ms. Miller: Yes. Thank you. Diana Miller from Industry. So, I like the way the question is phrased and I'm going to start from there and piggyback on what you said. How we tie the risk with the existing intended uses or new applications for existing uses. And I want to describe a little bit how we do in the industry, how we judge the risk and the importance of intended use for the medical device. So, even highlighted by the presentations you heard, think about the controls and existing controls you already have in place for an existing use and think maybe, you start from there, a more traditional-- An existing use, an existing device, and maybe you incorporate a new technology into it. What I want to point out is that a lot of times, I think the majority of times in the industry, when we think of risk and we think of controls, we think of the patient's risk and intended use and not target specific to the technology. So, you start-- And I'm going to describe very short the process. You start from assessing the failure modes for that particular device. So, I have this device, this is a use, what's a failure mode and what harm can you do, related to that intended use? So, that's agnostic of the technology and a lot of times what we heard today was close to that, but a lot of times we can apply this

kind of framework, and I think the first presentation was very close with what we do in the industry with red, yellow, and green. We do the same thing and assess and score the risk and the harm. That can be applied to this technology as well. So, after you evaluate the clinical implication of that harm and potential failures, then what do you do? Do you add special controls to mitigate them? Okay. So, you say for this particular harm, for this particular failure mode, I'm going to do this control, and this is going to mitigate that harm. And then after that, after you add all those specific controls, at the end of it all, you have a lady we call residual risk. After all of these controls and all the mitigations I put in place, what are my residual risks that I still need to mitigate? So, this is a very simple framework that-- I mean, it's not simple. You saw it's very close to what we heard today, but this is an existing framework we use, and we focus on the clinical harm first and it transcends the technology. I don't know if that's a word. I might have made up the word, but I think it does exist. But it transcends the technology. Thank you.

Dr. Bhatt: Great. Thank you. Dr. Jackson. And if you're done speaking, if you put your placard down, that would help me. Thank you.

Dr. Jackson: Thank you. One of the things that I was thinking about is-- I was trying to synthesize; it's so much data listening to everyone and everything that we read today. It relates to controls related to governance. And I'm starting to wonder if we need to think about having different governance controls because, to Dr. Shah's points, there's different GenAI models that will be used in medical devices. And even thinking about-- In the documents the FDA gave us, all the uses are not always going to be an image that goes in. Sometimes, if we have user interface, there's going to be a different interpretation that somebody is putting in as a part of the human-in-the-loop in different things. And I think having governance that is the same across both, I don't know that

we're going to have a solution, but if there's an image that's just going in, that's very easy to say, "This is a fact," but if there's a part of it that has human-in-the-loop, but there's still some interpretation and we're saying that has the same governance, I think we get into how do we determine which one is valid and reliable and all those other things that we were talking about? And so, I think that something that needs to be considered is having different governance controls depending on intended use and maybe even the model that is being used.

Dr. Bhatt: Dr. Radman.

Dr. Radman: So, I'm reading this question, and it has this point about real-world performance evaluation, and it leads me to want to just say that the risk of misdiagnosis is not new to generative AI. That's something we live with already with traditional AI, and that risk's getting more likely as the model drifts is also not new with generative AI. But I think-- As I'm building my framework of the collective thinking here, I'm realizing that the new challenge with generative AI is that the-- Is the exponential expansion of the input and output space of the data and the responses. And so, I'm not sure it's actually possible to adequately test in a satisfactory way in a premarket setting everything that could occur in that exponential input and output space. So, that really increases the importance in my estimation of real-world performance evaluation. And I think it is-- A new kind of regulatory framework is going to be needed. And that brings us back. We already talked about quality assurance labs and the day started with Dr. Califf talking about learning health systems, and I think he was right on it when he said-- He brought up learning health systems and he said he's not sure in his searches if there is a learning health system that can adequately validate an algorithm in the real-world setting. But I think that's where we need to go. And anything in the FDA's-- As someone put it, their gravitas of requiring, for example, closely related to the

predetermined change control plans, a plan to continue to monitor. And then, what happens when you detect you've drifted too much or there's new things that you didn't expect coming up, like we're discovering with other large language models giving us crazy responses. Every day we hear about new things. So, I'm just 'plus-one-ing' the real-world performance evaluation. I hope we could go further with it.

Dr. Bhatt: Dr. Shah, Dr. Maddox, and then questions-- Clarifying questions from the FDA, please.

Dr. Shah: So, I'm going to respond specifically to the part about risk and governance, which is on the screen Aubrey put. I think at the FDA, we have a very good-- Very good recall, market watch and safety alert system for drugs. Right? And that seems to be working pretty well. So, one would be for the FDA internally to think what could be a recall mechanism, what could be a safety alert mechanism and what could be a market watch, like notifications for people who are deploying these algorithms, to protect patients and avoid risk? And then, the FDA also launched a sentinel initiative, which I'm sure many of you are familiar with during COVID, which actually allowed reverse data collection from the field from patients. So, integrating some of those existing mechanisms, maybe not as regulatory heavy for some of these things, but having those safety alerts, etcetera, market watch and sentinel, should probably be in the framework of risk mitigation if things go wrong eventually, unfortunately.

And one point I want to bring, which is in my opinion not related to this question, is that Dr. Dreyer mentioned about CPT codes that there are only five CMS providers underwriting insurance for patients for these algorithms. I think that's a very big point that eventually patient-centered care is not going to work if you don't have providers-- If you do have providers being compensated by insurance companies. So,

then really what is the delta about helping patients get better care if their insurance providers are not going to pay for these tests? So, that can be tabled for now, but something to be considered. Thanks.

Dr. Maddox: Tom Maddox from WashU. I don't know if this made it to the list, but just to make sure, output variability to me is such a unique aspect of these tools, particularly when it's having a conversation with either a clinician or patient. And I just think we're going to have to think about how we both assess and control for that. I'm looking at that final phrase of real-world performance evaluation and I think there's probably a role in postmarketing evaluation as well, which we'll talk about tomorrow. Thank you.

Dr. Bhatt: Okay, really quick here, really quick there. And then FDA. No more cards up, please. Thank you.

Dr. Radman: I just wanted to add-- I think Tom is getting-- Is riffing on the point I was making about real-world performance evaluation as well. And I think the output variability increases the cost and burden of premarket evaluation. That's the issue here. There's just too much to evaluate. So, it really does necessitate the real-world evaluation to eventually do it thoroughly if you continuously monitor once it's deployed. And we could only do the best we can do in the premarket space, but just really resource constraint now that there's just so many outputs.

Dr. Botsis: Taxiarchis Botsis. A very brief comment inspired a little bit by Ms. Miller's comment about the failure mode. Obviously, we need to know the failures of the GenAI methodology. Absolutely. I wonder though whether it might be easier to specify and define a safe mode for those technologies. Just to give you an example, and imaging is not my expertise honestly, but if we have specific characteristics, image characteristics that may specify the generation of a text report for that particular image,

a CT scan for example, whether that set of features really take us to a safe mode for that particular technology.

Dr. Bhatt: To our FDA colleagues for clarifying questions. And then I'll summarize.

Mr. Tazbaz: Thank you so much. So, we have one clarifying question around when the Committee is talking about clinical validation, if we could just maybe describe what you mean by clinical validation, see if we're actually aligning from a regulatory language perspective. And then, we have a couple more questions around. Do you see any difference between a generative AI-- From a risk perspective or risk management perspective, a generative AI tool that is patient-facing versus clinician-facing? That's an important part of the equation of risk management. And then, there were some conversations about accountability. So, when something does go wrong, who's accountable for it? Right? Is it the developer? Is it the hospital system? Is it the individual clinician? And so, risk management kind of takes a different consideration there. And this is really not maybe pertaining to this conversation immediately, but I also would like you all to think about this concept of risk management and the complexity that we just saw, and even Dr. Jackson talked about multiple risk management frameworks or governance frameworks. And so, one of the things that we've been talking about is health equity and this is incredibly difficult to go implement. And so, how should we be thinking about this more collectively as an industry around this concept of risk management to ensure that these technology and tools can be used by everyone, not just the ones that have the infrastructure to be able to deploy and operate them? Thank you.

Dr. Bhatt: Chevon, may I ask you to comment on some of these?

Dr. Rariy: Yeah, absolutely. I was just trying to put my thoughts together, but I think obviously they're excellent questions around the risk management. I remember there were a few speakers who brought up the IRB process in your traditional clinical setting and it's more locally based. And I think we would benefit from considering something of that nature as well, not only a governance process, but some sort of regulatory oversight compliance, etcetera, that is locally based, and it will lend itself well to risk management in a similar way as the current state of risk management on the local level. So, I think there's benefit in having multiple layers, if you will, of risk mitigation strategies embedded within the-- In a systematic way. And part of that governance and risk mitigation strategy, perhaps the FDA can lend itself to creating at least an overarching scaffold of a framework. And that framework could include high level-- We saw the seven areas including equity as one, but at least providing various localities with the opportunity to adopt this framework and expound on it I think will help with accountability as well at the local level and perhaps provide more rigor in clinical care.

Secondly, I think it goes to your question, Mr. Tazbaz, around the patient-facing or direct to patient-facing generative AI versus direct to clinician. I think the key consideration here, as we heard it from some others, is that if we are reliant on expert interpretation for whatever medical specific use case of the generative AI output, then we need to take that into consideration that as a lay patient, they will not have that expert medical consideration. And so, they are two very distinct use cases that likely require two very distinct considerations.

Dr. Bhatt: Thank you. Going back to your patient-facing versus clinician-facing question, I just want to echo what Dr. Clarkson said just a little while ago. At the end of the day, does it lead to patient harm? I think sometimes we carry the assumption that

something that is clinician-facing may not directly harm the patient, and that's one of the risks we run in that terminology. And so, I think I would just make sure that we extend that to how does it affect the patient, whether the device is facing a patient or a clinician? Let's do Dr. Radman. Dr. Khubchandani, it would be great to hear from you after Dr. Radman, and then Dr. Jackson.

Dr. Radman: I just have a simple comment about the patient-facing that-- And I guess we were talking about-- I think Peter mentioned maybe we need a model for this group or this case and so forth. And so, we should also consider, while we want to protect patients and everything, there's also a case to be made for the benefits that patients may get, like consider a rural population that lives two hours away from the closest specialist, and whereas a generative AI can be a surrogate for that. So, there's a balance between the risks and the benefits. And the benefits can be far-- Increase for just thinking this particular rural populations, but I'm sure we could think of other use cases like that, whereas if you compare a benefit where a hospital is rolling out AI to cut costs and become more efficient or something like that, that's great, but it may be a lower benefit level there and therefore the risks should be accounted for accordingly.

Dr. Bhatt: That's great. Thank you. I'll piggyback off that for a minute and then go to Dr. Khubchandani. When thinking about the health equity part of your question, Troy, one challenge we face, because generative AI is oftentimes related to clinical guidelines or information or clinical decision support, is what the gold standard is. And when we think about whether or not we're delivering gold-standard clinical guideline-derived treatment throughout the United States throughout all of the different specialties, the answer is generally no. And so, it's how close does generative AI bring us to the gold standard versus how much better will the care get versus what it is now?

And so, just based on what Dr. Radman said, I think that's an important differential as well for us to think about.

Dr. Khubchandani and then, Dr. Jackson. And Troy, we haven't forgotten that you asked about clinical validity and accountability.

Dr. Khubchandani: Yeah. I think from the public health perspective, risk management-- When we submit IRB applications to do a study, we take the liability, the university takes the liability. Same for the healthcare system. They had a choice to adopt an AI technology or not to adopt it, so they would be in a way liable. My other comment is regarding how I see generative AI. It's like-- As Dr. Clarkson said, how do you define risk? So, I bought a wearable device. I did not lose weight. Risk. I lost some money. I could have invested it in a gym. But it kept misinforming me about my heart rate and I died. Now, that's a different type of risk. And so, risk has to be defined. Risk assessment has to be defined and tied up with risk management. That's my comment here. Thank you.

Dr. Bhatt: Dr. Jackson.

Dr. Jackson: I wanted to speak to the accountability question. I think it's a really interesting question, and I'll be honest, I'm not sure that it's necessarily fully under the FDA's purview, right? There's an article in the New England Journal of Medicine from, I think, January, February of this year, and they were looking at generative AI and accountability and liability for providers, and they laid out multiple legal cases of when software's medical devices come up and who ended up being accountable and even liable. And I think risk management and liability are two different things, but that's what they were talking about. How do we differentiate between the two? And it showed that every time there was a different answer, depending on what was happening. And so, I don't know that we can have a definitive accountability in this case, but thinking

about it from a health equity lens, which for me has been very important in thinking about all these discussions. I wonder if we could borrow from some of the value-based care work that has been done in thinking about shared risk management and what that looks like when a healthcare system or a practice or a clinician decides to deploy GenAI in their work, and how do we look at shared risk management as a part of that that incentivizes all parties to make sure that we're looking at safety and quality making decisions? And I think that is a definite pathway to also improving health equity if everyone is ensuring that. Because I think one of the things that could happen is people will decide, if the risk management is too high, not to deploy some of these tools, which would increase disparities because they feel like they have too much accountability.

Dr. Bhatt: Thank you. Dr. Kukafka.

Dr. Kukafka: I just wanted to talk a little bit more about the patient-facing that you brought up in the context of risk mitigation. I don't know if the FDA considers exasperating health disparities as a risk or is that out of context, out of your purview? I mean, my point is if we do patient-facing, we have to design for different levels of literacy and numeracy to be-- So that we don't exasperate disparities. Interesting. I just completed a study where patients were able to download their own EHR data, review their data, and we found that the patients with lower literacy levels, lower education levels were more likely to do that than patients more educated. Now, we have hypotheses that they might be more-- Higher level of information-seeking because they do have issues with obtaining and interpreting health information. But again, is that a risk for the FDA to consider exasperation of health disparities? That's a question. And then, if the answer is yes, then the usability has to address that.

Dr. Bhatt: I think that's an incredibly helpful comment. And I would say I think it's essential for us as a Digital Health Advisory Committee, not just today, for our other

sessions as well, to think about digital and health literacy for the patient-facing parts of whatever we are going to look at, whether it's generative AI or something else in the future. So, thank you for bringing that up, and I think it dovetails nicely with what Dr. Jackson was saying as well about exacerbating the discrepancy in access or usability, and we can probably break those down as we think about this further. Dr. Elkin.

Dr. Elkin: Yeah. I'd just love to build on that. It touched something in my heart as a clinician. Many years ago, when people started getting information on the Internet, they'd come into your office, and some of my colleagues really didn't like that, but I loved it because it was getting involved with their own healthcare. And the more they were involved in their healthcare, the more we could have a great conversation and draw them in and try to help solve the problems that they're coming in with. And it turned out that, at least from my practice, this was very helpful, even if the information that they were worried about didn't really apply to them. I could explain why and alleviate fears and tell them when they needed to come in and encourage them to continue to be involved in their own healthcare. And I see the ubiquitous nature of these devices where they can be on your phone. Right? It doesn't have to be a very expensive computer that you have. It's going over the Internet and the main hardware that's running is in the cloud. So, it could give people that don't have access to high quality medical information much more medical information for them to bring to their clinicians to get involved in their care and proactively, I think that will help to increase the quality of care and decrease disparities.

Dr. Bhatt: Okay. Mr. Radman, maybe last comment. I'm going to look over at Dr. Clarkson and Diana Miller for any final comments they may have. And then I will read to you our concerns, but not our summary in the interest of time, Mr. Tazbaz.

Dr. Radman: Thank you. Regarding accountability, I think there should be strong requirements for labeling studies because, for example, chat GPT already has a disclaimer that it's not a medical device, but we've already seen data that shows its adoption and use by clinicians already as a second opinion. And we all know how ineffective the listing of the side effects on drug commercials can be at the end of those commercials. So, I think if-- We discussed already the importance of having some kind of label saying, "This output is generated by a generative AI or large language model," and maybe even saying-- I've seen people recommending-- "Which has a propensity to hallucinate." Perhaps even we could have a premarket evaluation to determine hallucinations occur X percentage of the time or something like that. But do people understand that label? Whatever the label says. So, that's really important to verify with an adequate study with a diverse user group that reflects who's going to be using the device.

Dr. Bhatt: Great, thank you. Let's do a rapid fire. Just a quick round. Does anybody have a 30-second answer to clinical validation? Who used that phrase? And do you want to claim that you didn't explain to Troy what you meant? No?

Dr. Maddox: I actually think it came up in Dr. Dreyer's presentation. So, first of all, can you just expound on your question real quick to make sure we're answering it?

Dr. Diamond: Dr. Dreyer, I think you were saying that you didn't see a lot of value in some types of comparator studies, and instead we're advocating for a different type of clinical validation. Could you explain?

Dr. Dreyer: Yeah. I'll talk about this in the open part tomorrow to talk about monitoring as well where I'll get to this. But if you think of software testing, just straight software testing, the next step would be you would do a performance test on that software. That means that you have data, but you don't have experts. The next thing

you would want to try and do is consider doing a comparative test with experts and the data to say with and without the AI. Next, as we deploy every time we deploy AI, we do clinical validation. What I mean by that is think of it as a one-time monitoring. So, we deploy this in a mode where we can see the exact output of what's coming at the facility using our data and our experts with the algorithm. So that's the distinction between a comparative effectiveness of premarket testing with a pre-deployment testing at a facility. And then, as that continues on, either intermittent or continuous, that's the monitoring process. So, what I mean by clinical validation or validation of the test is that it's validated on that site to prove that it is at a level of acceptability at that site. And I make that distinction also because what we've seen continuously with the previous models, and I think it's going to be the same here, is that because this works at five sites, it doesn't mean it'll work at the next five. And then the question that comes, as we've discussed, is if it's working at 50 sites but it's not working at 10 sites, is it recalled or is it not? Is that a good thing or a bad thing? And so, I don't think you can premarket decide that. I think you have to validate at each facility to be able to make that distinction.

Dr. Diamond: So, you're talking about site-specific validation in the context of real-world use.

Dr. Dreyer: Or even pre-real-world use, but essentially real-world evidence by not maybe the clinical deployment of it, but the preclinical deployment of it to test it with site users, site patient data and the algorithm.

Dr. Bhatt: Any last 30-second comments from the Panel? Ms. Miller.

Ms. Miller: I want to address a little bit about the patient-facing and the access to healthcare. And I think there are two important aspects to think about when we design a medical device targeting patients. You obviously have to think of different personas and

who do you design the device for? And I think it's a real benefit, and that's what we mean by intended use population target. But a real good benefit that we still need to be mindful of is the access to healthcare. It's increased ever since phone's been around. There are medical devices, applications on the phone that contribute to a lot of benefits to the patient's access. And even though maybe we don't get to try it first-- We keep iterating onto it. And I think the FDA acknowledges this with the PCCP framework and all of that, that allows faster iterations on these devices because something we need to think about when we design for patients is that we need to have a consumer mindset. And in a medical device industry, we're a little handicapped in a way that we can't really act that fast like tech companies and we cannot deploy that fast our medical device apps to those devices and technologies. So, we really appreciate the PCCP, and we need to realize that we need to give the same experience like the Apples of the world-- I don't want to call names, but the big tech companies. We need to be on the same bar and that's what it's hard to strive for.

Dr. Bhatt: Okay. Thank you so much. So, in the interest of time, I have the documentation here of the summary that I will hand in for the record. And I will leave us just with some concerns that the Panel has according to this question that include the following. This is a new way of presenting information that seems more humanistic and gives the impression of real intelligence. That is concern number one. Number two is, whereas we are thinking about national guidance, local governance may be most important and value sets appear to be local. The third is clinicians are busy and may not engage in training the way we are expecting. So, our reliance on training to ensure that generative AI systems work may not work as intended and it may be challenged. The risk of patient harm must be central to this discussion. There's an increased risk of lack of accountability in healthcare with the addition of generative AI to care delivery.

Governance may need to differ across different types of generative AI. This may be differentiated across type of input, type of model and potential clinical outcomes. And there's a suggestion that we could perhaps use infrastructure for drugs and other devices that already exist at the FDA and expand them to generative AI-related devices. Digital health and literacy is essential to include in risk evaluation, monitoring and mitigation if we are to keep health equity central to our efforts. And although accountability may not be under FDA purview, responsibility may lie with health systems that deploy GenAI devices or the companies that create them and others at this time, because there is not a large body of clear precedents. Shared risk management may likely be the best concept for now as it will incentivize all parties to look at risk mitigation as their responsibility.

Mr. Tazbaz, is this an adequate summary for now? Thank you.

Day 1 Closing Remarks

Dr. Bhatt: With that, I would like to thank the Committee and the FDA for their contributions. I would also like to thank again the open public hearing speakers, patient, industry, healthcare provider, academia and FDA for their remarks.

And before we adjourn, I want to ask the FDA staff seated at the table if they have any final comments.

Mr. Tazbaz: Thank you, Dr. Bhatt. I would like to echo the same gratitude towards the presenters, the public commentary and of course the Committee itself, but also to you, Dr. Bhatt, for running a very tight and productive session. So, we're looking forward to seeing-- We've been able to capture several action items that could come out of it, but we're looking to collaborate further to drive more of these action items that we might be able to tackle collectively. So, thank you.

Adjournment

Dr. Bhatt: Great. Thank you. The November 20th session of the Digital Health Advisory Committee is now adjourned. We'll continue this agenda tomorrow at 9:00 a.m. sharp. Have a wonderful night and thank you.