U.S. Food and Drug Administration
Center for Devices and Radiological Health

Digital Health Advisory Committee (DHAC) Meeting on the topic
"Total Product Lifecycle Considerations for Generative AI-Enabled Devices"
Summary Minutes for Day 1 - November 20, 2024

Introduction

On November 20, 2024, the FDA's Digital Health Advisory Committee, or DHAC, met to discuss and provide recommendations on how the use of generative AI (GenAI) may impact safety and effectiveness of medical devices enabled with this technology. This was the first meeting of the Advisory Committee that was recently established to address digital health technologies. The topics under discussion included premarket performance evaluation, risk management, and postmarket performance monitoring for GenAI-enabled devices.

Dr. Ami Bhatt, Chair, noted for the record that members constituted a quorum, and proceeded with the introductions of the members identified in the meeting roster.

Conflict of Interest Statement

Mr. James Swink, Designated Federal Officer for the DHAC, read the Conflict of Interest Statement and stated that all members and consultants of the Committee are subject to federal conflict of interest laws and regulations, and that the FDA has determined that all of them are in compliance with such laws. No conflict-of-interest waivers were issued.

Dr. Robert Califf, FDA's Commissioner, delivered welcoming remarks emphasizing the importance of collaboration and innovation in advancing digital health technologies, particularly AI. He highlighted the Advisory Committee's role in providing expert advice and diverse perspectives to support the FDA's mission. Dr. Califf pointed out AI's potential to tackle healthcare challenges, improve life expectancy, and reduce health disparities. He stressed the need for real-world validation systems for AI in healthcare and expressed concern about prioritizing financial returns over patient outcomes. He concluded by urging the Committee to consider solutions for ensuring the safe, effective, and equitable use of AI technologies.

Dr. Michelle Tarver, Director of the Center for Devices and Radiological Health, delivered the opening remarks, emphasizing the FDA's commitment to patient safety and timely access to innovative AI-enabled medical devices. She highlighted the potential of these technologies to improve healthcare, particularly for underserved populations in rural and low-resource areas. Dr. Tarver stressed the importance of ensuring these technologies are safe,

effective, and accessible. She also called for proactive, inclusive discussions to advance AI equitably and ethically. She concluded by setting the stage for discussions on generative AI and the total product lifecycle of AI-enabled technologies, focusing on their benefits, risks, and clinical outcomes.

FDA Perspective - GenAI in Medical Devices

Mr. Troy Tazbaz commented on the growing significance of digital health technologies and highlighted DHAC's role as the FDA's first Advisory Committee focused on crosscutting digital health issues, aiming to guide regulatory considerations for a wide spectrum of medical devices. He underscored the transformative potential of GenAI in healthcare, noting its ability to produce new outputs from vast data, which offers significant opportunities but also introduces unique regulatory and operational challenges. Unlike traditional AI models, GenAI's ability to generate novel outputs raises complexities in ensuring the safety and effectiveness of medical devices.

To address these challenges, Mr. Tazbaz advocated for a total product lifecycle approach, considering safety and performance from design through postmarket monitoring. He outlined the agenda, which includes discussions on premarket evaluation, risk management, and postmarket performance monitoring of GenAI.

He expressed confidence that the discussions would help the FDA balance innovation with safety, fostering trust in GenAI technologies. He concluded by previewing FDA's upcoming presentations on regulatory science challenges and strategies for managing changes in AI-enabled devices.

Sub-topic: Premarket Performance Evaluation

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

Dr. Aldo Badano highlighted the evolving role of AI and machine learning, particularly GenAI, in healthcare and medical devices. He discussed the wide range of applications for GenAI-enabled devices in healthcare, emphasizing the balance between their potential benefits in performance and productivity and the challenges they present, especially in terms of regulation. Some key challenges he noted, include defining the scope of products, oversight of adaptive systems, handling errors like hallucinations in AI models, and ensuring data diversity and performance monitoring to prevent biases. He also touched on the importance of transparency in AI tools to build trust with users.

To address these challenges, Dr. Badano outlined three evaluation strategies for GenAI technologies: benchmarking, expert evaluation, and model-based evaluation. Benchmarking involves testing models against standardized datasets, which allows for large-scale comparisons but risks overfitting to specific data distributions. Expert evaluation, while adaptable and clinically relevant, is resource-intensive and subjective. Model-based evaluation uses automated tools to assess performance and benefits from careful oversight to avoid issues like inter-model leakage. Dr. Badano's team at CDRH is actively working on model-based evaluations for radiological impressions generated by GenAI, focusing on factual accuracy and performance metrics.

To conclude, Dr. Badano emphasized that device performance evaluation considers the intended use and associated risks, and that both premarket and postmarket evidence will be crucial as GenAI technology evolves. He called for continued collaboration to develop a balanced approach to ensure the safety and effectiveness of these rapidly advancing technologies.

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

Dr. Faisal Mahmood discussed the evolution of AI models in computational pathology, focusing on the transition from supervised learning to self-supervised learning, and the introduction of GenAI. Initially, AI models relied on labeled data for supervised learning, but self-supervised learning allowed for the development of foundation models that could learn from pathology images with minimal labeled data. These models, which integrated modalities like pathology images and immunohistochemistry (IHC) data, improved tasks such as IHC quantification and survival predictions.

Dr. Mahmood's team further advanced this by integrating molecular data, like transcriptomic data from next-generation sequencing, with pathology images. This integration was particularly useful for detecting rare diseases and predicting treatment responses, especially in tertiary medical centers with limited treatment-response data. The combination of molecular and histological data helped develop more robust models for predicting outcomes and identifying biomarkers.

The team then explored GenAI for creating a unified model encompassing human pathology knowledge. Despite challenges like the lack of region-level annotations on pathology slides, they developed, a GenAI tool that could answer pathology-related questions, generate diagnostic reports, and suggest additional tests. The tool was successfully tested in resource-limited settings, where it generated pathology reports from cell-phone microscope images.

The team also demonstrated AI agents in biomedical research, such as agents that autonomously handled tasks like training models to predict treatment responders and non-responders. These agents could plan, segment tissues, extract features, and even generate code, marking progress toward autonomous discovery agents.

Concerns about fairness and bias in AI were addressed, particularly regarding demographic data and imaging variations. The team used contrastive learning and bias mitigation techniques to ensure fairness across diverse populations and clinical settings.

Finally, Dr. Mahmood stressed the importance of evaluating GenAI models like the tool presented. He described how quantitative and qualitative assessments from pathologists were used to evaluate the models' clinical relevance and accuracy, with ongoing refinement necessary to ensure these models effectively link diagnostic descriptions to clinical outcomes.

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

Dr. Parminder Bhatia discussed the transformative potential of GenAI in healthcare, emphasizing both technological advancements and the need for responsible implementation. He introduced foundation models, which are multimodal AI systems capable of processing and synthesizing diverse data types, making them highly adaptable for various healthcare applications. These models can be fine-tuned for specific tasks like analyzing medical images or generating reports, significantly improving workflows in areas like radiology.

He highlighted examples of foundation models in healthcare, such as their use in breast cancer care pathways, where they streamline processes from screening to diagnosis and treatment. By synthesizing data from different clinical modalities, these models can enhance the efficiency and accuracy of medical workflows, particularly in complex fields like oncology. Dr. Bhatia also noted their potential to quickly adapt to new areas, accelerating the development of care pathways for emerging diseases or treatments.

Dr. Bhatia stressed the importance of responsible AI implementation by establishing frameworks to manage risks. He outlined four key pillars for safe rollout: defining the intended use of AI models, adopting robust risk management practices, conducting rigorous premarket evaluation, and ensuring effective lifecycle management. He emphasized the need for clarity in AI use cases to ensure safety and regulatory compliance. Risk management can anticipate issues like hallucinations (AI generating incorrect information) and incorporate strategies such as human oversight, reasoning algorithms, and visual grounding techniques to improve diagnostic accuracy.

To conclude, Dr. Bhatia expressed optimism about the potential of GenAI and foundation models to revolutionize healthcare. He noted that with proper controls, these technologies can be integrated into medical devices to enhance functionality, improve patient care, and ensure safety and reliability. He highlighted how predetermined change control plans (PCCP) allow for quick adaptation to new advancements without compromising regulatory standards.

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

Dr. Pranav Rajpurkar focused on evaluating GenAI applications in clinical settings, particularly for generating and assessing medical reports. He discussed the significant potential of AI in summarizing clinical notes, generating reports from images, and enhancing image analysis, driven by multimodal data and natural language processing. A key challenge he addressed was the evaluation of AI-generated reports, as traditional metrics like BLEU and BERTScore fail to capture medical context.

To improve evaluation, Dr. Rajpurkar's team developed new metrics to better evaluate medical accuracy, anatomical context, and negation. These models were further refined using HeadCT-One, which compares AI-generated reports to expert ones based on anatomical and medical terms. He emphasized the importance of expert evaluation and proposed RadCliQ, a composite metric to improve alignment with expert ratings, and FineRadScore, which identifies specific errors in AI reports based on clinical significance.

Dr. Rajpurkar also discussed the development of MedVersa, a generalist model that generates reports and performs tasks across various modalities, outperforming specialized models. He highlighted the importance of human-centered evaluation, noting that expert-written reports are often preferred, especially for abnormal cases, though AI-generated reports saved time and boosted confidence in radiologists.

He pointed out the limitations of current benchmarks for large language models (LLMs) in medicine, particularly the gap between medical exam evaluations and real-world patient interactions. He also highlighted an AI tool helping patients understand medical scans by linking reports to image regions, illustrating GenAI's potential to empower patients.

To conclude, Dr. Rajpurkar emphasized the challenges and promise of GenAI in healthcare, stressing the need for human-centered evaluations and continued research to ensure safe, effective integration into medical practice.

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

Dr. Apurv Soni asked about the role of AI assurance labs in evaluating GenAI, particularly in relation to FDA metrics.

Dr. Pratik Shah inquired about regulatory pathways for GenAI models not fully trained on clinical data, suggesting two approaches: narrowing the model scope or shifting from traditional metrics to preference testing. Dr. Rajpurkar agreed on defining intended use based on real-world utilization, considering elements such as specific user groups.

Dr. Peter Elkin proposed the use of a model card documenting training datasets and model properties, along with monitoring rare errors and model drift. He emphasized the role of clinical informaticians in ensuring safe use. Dr. Bhatt highlighted the opportunity for postmarket surveillance to understand real-world usage.

Dr. Rita Kukafka raised concerns about clinical endpoints for GenAI evaluations, emphasizing the importance of outcome measures such as clinician or patient preference. Dr. Bhatt agreed on focusing on structure, process, and outcomes for GenAI evaluations.

Dr. Thomas Radman discussed preference testing and sample sizes for trials involving GenAI devices. Dr. Rajpurkar outlined a trial design comparing AI-generated drafts with standard care, focusing on clinical significance and error rates. He noted challenges related to user diversity.

Ms. Diana Miller asked several targeted questions regarding data. Dr. Mahmood explained the data diversity used in his team's pre-trained model. Dr. Bhatia discussed the selection and adaptation of foundation models, including examples of cross-domain applications.

Dr. Rajpurkar described evaluations of GenAI in 12 specialties, including diagnostic accuracy and preference testing. Dr. Laura Stanley raised concerns about trust in AI systems, with Dr. Bhatia emphasizing the importance of explainability and chain-of-thought reasoning to build trust.

Dr. Chevon Rariy questioned Dr. Mahmood about end-user information and labeling in GenAI. Dr. Mahmood acknowledged challenges in performance quantification and traceability. Dr. Maddox raised the issue of managing automation bias, and Dr. Bhatia suggested a phased approach for auditing and monitoring AI outputs.

Dr. Rariy asked about assessing model maturity for clinical decision-making. Dr. Bhatia emphasized that maturity involves evaluating post-deployment metrics, external validation, and continuous monitoring.

Open Public Hearing

Dr. Noah Zimmerman discussed the potential of GenAI in healthcare, particularly for documentation and diagnostics. He stressed the importance of regulatory frameworks to ensure safety while fostering innovation. He also noted that not all GenAI uses in healthcare are medical devices and that existing FDA frameworks could evaluate technologies based on intended use and risks. Dr. Zimmerman proposed modernizing the FDA's MAUDE database to enhance post-market surveillance, tracking the evolving risks and applications of GenAI.

Dr. Bernardo Bizzo presented on the American College of Radiology (ACR) Assess-AI program, which connects over 40,000 imaging entities nationwide. He highlighted the ARCH-AI program, the first national AI quality assurance initiative for radiology, and discussed the performance monitoring of AI tools through Assess-AI. He noted that, the program compares AI model performance at individual sites with national benchmarks. Current data is being collected from 15 pilot sites, with AI vendors focusing on clinical use cases like intracranial hemorrhage and pulmonary embolism on CT scans. Dr. Bizzo concluded by emphasizing the collaborative work of the ACR with various healthcare stakeholders to ensure the real-world effectiveness of AI tools, particularly for uses like drafting radiology reports.

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

Dr. Bhatt led a discussion with Dr. Bizzo on monitoring safety and efficacy in GenAI applications in radiology. Dr. Bizzo explained that GenAI's probabilistic nature, which produces varying outputs from the same data, necessitates focus on the content of AI-generated outputs and their impact on patient safety and healthcare provider decisions. He highlighted ongoing efforts to manage variables influencing care management and safety.

Dr. Bhatt asked about specific safety metrics, and Dr. Bizzo emphasized monitoring changes between draft and final reports, particularly clinically meaningful adjustments. This work is still in early stages. Dr. Botsis inquired about metrics for different use cases, and Dr. Bizzo confirmed they would be tailored to each AI model, with local validation and continuous monitoring.

Dr. Taxiarchis Botsis also asked about thresholds for model performance. Dr. Bizzo explained that performance would be rated by clinical skill level and that unsafe reports would be excluded until they met thresholds. He also noted regulatory considerations in setting these thresholds.

Dr. Elkin raised concerns about bias in training data, particularly regarding variations in normal anatomy across races, ethnicities, and age groups. Dr. Bizzo acknowledged the challenge

and emphasized the importance of healthcare professional validation to distinguish normal anatomical variations from pathological findings.

The discussion continued with Dr. Radman questioning whether GenAI should have unbounded outputs or more narrow focuses. Dr. Zimmerman explained that both approaches are feasible, depending on the intended use of the AI system.

Dr. Soni suggested that the FDA could serve as a clearinghouse for collecting and analyzing data on the real-world use of GenAI tools. Dr. Zimmerman agreed, proposing a system like the Adverse Event Reporting System for drugs to track GenAI use and ensure post-market surveillance.

Dr. Bhatt raised concerns about error reporting in the industry, and Dr. Zimmerman pointed out that manufacturers could integrate APIs into their devices to streamline reporting. Dr. Jessica Jackson raised concerns about data integrity, specifically regarding the potential manipulation of medical images. Dr. Bizzo reassured that AI data comes from real clinical workflows, following existing regulatory guidelines to ensure data authenticity.

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

Ms. Aubrey Shick opened the discussion by outlining the FDA's request for input on premarket performance evaluation of GenAI-enabled devices. She emphasized concerns about the limited data on pre-trained generative models and the dynamic, ever-evolving nature of foundation models. She introduced the 4 sub-questions to be discussed by the Committee.

Question 1. *Premarket Performance Evaluation: Please discuss what specific information related to generative AI should be available to FDA to evaluate the safety and effectiveness of Gen AI-enabled devices considering, for example, that foundation models leveraged by the Gen AI-enabled device will change over time and that there may be limited information available on the training data utilized for these pretrained generative models.*

Question 1a: *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? What information is particularly valuable to evaluate the safety and effectiveness for devices enabled with generative AI in comparison to non-generative AI?*

As posted in the meeting 24 hour Summary, available at: https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-20-21-2024-digital-health-advisory-committee-meeting-announcement-11202024#event-materials

The Committee generally agreed that the device's description or characterization for GenAI-enabled devices should include information on the device's intended use and indications for use including a detailed description of specific use cases and the intended care environment.

They also emphasized the need to include whether there is the intention for a human to be in the loop, a description of the AI-human interactions, and the expertise needed for the user including any training. The Committee agreed on the importance of providing information on the data used to develop and test the device, such as the dataset size, types, and patient demographics (whether it includes a diverse dataset and diverse locations of care). The Committee also recommended providing information on the foundation models the device relies upon, such as guardrails or constraints imposed on the device's input and output, known or potential failure modes of the device, the adaptivity of the device (e.g., whether the models used are "locked," etc.). They also recommended providing information on specifics regarding cybersecurity and privacy and how the transparency of the device is supported. The Committee noted that a standard data sheet or model card could be a helpful approach to provide some of this information. The Committee also agreed that the other information for device evaluation be included for these devices, including, for example, information related to risk management, change management, and quality systems.

Question 1b: *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?*

As posted in the meeting 24 hour Summary, available at: https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-20-21-2024-digital-health-advisory-committee-meeting-announcement-11202024#event-materials

The Committee generally agreed that the device's premarket evaluation should include a characterization of the device's performance. The specific performance metrics would depend on the specific intended use of a device; for example, for a diagnostic device, it would generally include sensitivity and specificity as well as other metrics for diagnostic device performance. The Committee noted that the characterization of the device's performance should also include performance in different populations and settings, as applicable for the intended use of the device. The Committee recommended that the premarket evaluation include a characterization of the repeatability and reproducibility, as well as the uncertainty of measurement, including, for example, uncertainty estimates, hallucination rates, error rates, expert evaluation, severity of error, degree of corrective measures taken, and results related to stress-testing. The Committee noted that this may be difficult for sponsors to provide to FDA when the GenAI-enabled device uses a third-party foundation model. In such cases, the Committee stated that it may be important to provide information on the data that the third-party foundation model was trained on and/or tuned with, to the extent possible. The Committee stated premarket evaluation of a GenAI-enabled device could include benchmarking against other models. The Committee agreed that premarket evaluation should include a proposal for postmarket monitoring of the GenAI-enabled device and that this could be particularly important when more limited information is available on the foundation model it relies upon. Finally, the Committee noted that the types and level of information for premarket evaluation generally should be commensurate with the risk of the device, consistent with FDA's existing risk-based approach.

Question 1c: *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?*

As posted in the meeting 24 hour Summary, available at: https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-20-21-2024-digital-health-advisory-committee-meeting-announcement-11202024#event-materials

The Committee generally agreed that the user interface, and in particular, explanation of the device inputs and outputs (to the appropriate user, in the device's context of use), will be an important consideration in promoting transparency to the user and enabling trust in these devices. The Committee agreed, on the topic of transparency, that it may be important for users to know when they are using a GenAI-enabled device, and that it may be important for patients to know how a GenAI-enabled device contributed to their care. This could be done, for example, through use of a label that accompanies the generated output or information. It may also be important for users to know that such outputs may not necessarily be reproducible. Regarding device inputs, the Committee noted that the varying possible inputs for a GenAI-enabled device, such as text, images, or multimodal inputs, may not be a typical set of inputs for clinicians, patients or caregivers. As such, the Committee noted that it will be important to explain to users what information the device used for its decision making or other actions. The Committee discussed the importance of training, for clinicians, patients, and caregivers, so that users can understand how a device should and should not be used. The Committee noted that when GenAI-enabled devices are intended to be used by patients or caregivers, additional information may be needed.

Question 1d: *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. Examples of known metrics to support discussion may be modality-specific such as for generative text (perplexity, quantitative comparison to reference text), for generative images (Frechet 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.*

As posted in the meeting 24 hour Summary, available at: https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-20-21-2024-digital-health-advisory-committee-meeting-announcement-11202024#event-materials

As described in Question 1b, the Committee generally agreed that premarket evaluation should include a characterization of the device's performance, including, for example, sensitivity and specificity, or other performance metrics appropriate for the intended use of the device. The Committee noted that provided information may also include the established upper and lower

bounds for performance. To build on this, the Committee noted it could be helpful to specifically consider edge cases as part of testing, to help build an understanding of frequency and types of errors made by the GenAI-enabled device. In all instances, including those known metrics mentioned in the question, the Committee noted that communication of the results is important, and should include information about how the model reached its determination or output, when available. Finally, the Committee noted that data drift metrics will be important to help evaluate if models remain accurate as it pertains to safety and performance, highlighting the importance of continued monitoring of devices past the point of premarket evaluation. Separately, the Committee also noted that the device's safety and effectiveness may need to be assessed through distinct and new evaluations. Regardless of the metrics used, the Committee generally agreed that due to the uniqueness of each GenAI-enabled device, clear communication and explanation of a particular device's safety and performance metrics to FDA and users will be important.

Sub-Topic: Risk Management

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

Dr. Michael Schlosser stated that HCA Healthcare emphasizes the importance of governance and risk mitigation in AI use, including using a responsible AI framework. He described that the company also uses a human-in-the-loop strategy and stressed that all AI strategies are fundamentally data strategies. He explained that HCA Healthcare employs a risk-based approach to AI deployment, focusing on assessing risks and benefits through a framework with seven domains, such as safety, security, and transparency, and a detailed Risk Register to assess potential risks, including privacy, security, and hallucination. He explained that HCA considered the human-in-the-loop strategy to be essential for ensuring AI models' accuracy, especially in high-risk healthcare contexts and to prevent errors from affecting patient records.

Dr. Schlosser elaborated on six key concepts for designing effective human-in-the-loop systems. His concepts included promoting transparency, displaying levels of uncertainty, encouraging active participation, providing training, designing for trust calibration, and implementing feedback mechanisms. He shared an example of a nurse handoff tool, where users provided feedback on five critical variables in relation to the quality of the model using a scale: factuality, coverage, organization, conciseness, and helpfulness. In his example, by using these feedback loops, the system was iteratively improved.

In addition, Dr. Schlosser emphasized that, when evaluating GenAI systems, it is important to assess the entire system, including the human interaction and feedback, rather than focusing just on the technical performance of the model itself. He noted that for AI to be successful, it should be continuously used and refined. The speaker also highlighted the need for

continuous improvement and monitoring of AI models, as well as the importance of combining foundational models with context-specific data for success.

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

Dr. Keith Dreyer focused on the differences between narrow AI and GenAI, specifically within the context of risk determination, regulatory controls, and safe innovation in healthcare.

He began by describing the early use of AI in healthcare. He described how this innovation challenged regulatory frameworks, especially around whether AI should be classified as a medical device. He noted that, while AI has the potential to improve diagnostics, its adoption has been slow due to regulatory hurdles, particularly due to a lack of risk stratification. As a result, AI adoption has been slow in radiology.

He highlighted regulatory frameworks, particularly around the use of device software functions, and emphasized the need for clearer, more effective ways to evaluate AI tools.

The discussion then turned to the future of GenAI in healthcare, where the speaker emphasized the challenges in regulating this technology. He advocated for modifying current regulatory practices to incorporate pragmatic clinical evaluation and ongoing monitoring, as well as standalone performance testing, to test accuracy and better address the risks posed by GenAI tools.

Finally, Dr. Dreyer proposed considering a broader approach to AI regulation, including considering products that do not currently fall under the medical device definition. He pointed out the importance of monitoring and validating GenAI tools, especially in areas like diagnostic reporting, to allow for safer, more efficient innovation while managing the associated risks.

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

Dr. Danielle Bitterman discussed the role of large language models (LLMs) to enhance clinical care while managing risks to ensure safety. The speaker raised her concerns regarding their performance in the real-world as safe clinical applications that genuinely benefit patients and clinicians. She described how although LLMs have a lot of potential, they can introduce unique risks that can be proactively managed to balance innovation with safety.

She explained that LLMs undergo additional fine-tuning processes, such as instruction tuning and preference tuning, to improve their ability to follow specific instructions and preferences. However, she clarified that these tuning processes can also introduce new risks, such as models becoming overly compliant, which is particularly concerning in the context of

healthcare, where misinformation could have serious consequences. She also stressed the importance of transparency in training data to better understand a model's behavior and potential risks.

She discussed LLM risks for clinical applications and outlined various levels of risk, including systems privacy and security, patient safety and clinical effectiveness, workflow integration, and ethical and legal concerns. She noted that data governance, security protocols and user controls are critical for mitigating these risks in healthcare settings using LLMs. However, she also pointed out that current benchmarks do not reflect real-world clinical applications, and new methods may be needed for evaluating LLM performance. Additionally, she observed that integrating these models into clinical workflows raises new challenges, such as potential automation bias and over-reliance on LLMs by clinicians. She highlighted that ethical concerns, such as equity, transparency, and accountability, can also be addressed to build trust in LLMs.

Finally, Dr. Bitterman concluded by stating the importance of a measured approach to mitigate risks and ensure a safe and effective integration of LLMs into healthcare systems.

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

Dr. Gabriella Waters discussed the challenges and opportunities of GenAI in healthcare. She emphasized that LLMs are probabilistic and can make mistakes, which is why it is a challenge to deploy such models in high-risk settings like clinical environments. She addressed the importance of understanding their biases and potential for harm and highlighted the importance of model transparency and explainability, two separate but related concepts.

She discussed concerns around the need for robust training protocols and validation procedures, including testing, evaluation, validation, and verification (TEVV) processes. She stressed the importance of continuous feedback from users and real-world testing to identify potential risks and continuously improve the models.

Dr. Waters also highlighted the importance of post-market surveillance and performance benchmarking, noting that benchmarking tells us about capabilities and performance, but does not tell us about risk.

She concluded by mentioning the increasing adoption of GenAI into healthcare, from mobile devices to online doctor interactions, and emphasized the dangers of over-reliance on AI. She emphasized that human oversight and critical testing are essential to ensure AI enhances healthcare without compromising trust or safety.

Open Committee Discussion Q&A

After the presentations, Dr. Bhatt opened the Open Committee Discussion and allowed time for participants to ask clarifying questions.

Dr. Soni asked Dr. Dreyer to expand on how efficiency can be measured and monitored. Dr. Dreyer responded by saying that he believes it is the market's responsibility, not the FDA's, to measure efficiency.

Dr. Radman questioned whether it is valid to compare new devices to older ones using the FDA's substantial equivalence framework. Dr. Dreyer questioned whether equivalence tests are sufficient, especially in GenAI, and suggested validating devices on a case-by-case basis.

Ms. Miller asked Dr. Schlosser to clarify the intended use of the nurse handoff model. Dr. Schlosser clarified that the model is meant to facilitate the handoff of patients between nursing shifts, but it is not a medical device.

Dr. Kukafka asked about how to train clinicians and patients to mitigate the risks of AI models. Dr. Schlosser noted that training clinicians is crucial, and he believes those who fail to train may become obsolete. However, training patients, especially in diverse populations with different understandings, is much more difficult.

Finally, Dr. Elkin first asked if the presenters could share their risk assessment surveys. Dr. Schlosser responded that they're happy to share it and mentioned that they're collaborating with the FDA to study the framework's effectiveness. Then, Dr. Elkin asked a second question was about whether AI models can be tailored to specific populations to ensure better accuracy. Dr. Dreyer agreed that targeting specific populations could improve accuracy.

Committee Discussion of the FDA's Questions

Members of the Digital Health Advisory Committee were asked to respond to FDA Question 2 in relation to risk management and mitigation in the context of GenAI for medical devices. Specifically, the question asked about the new opportunities GenAI has enabled and what new controls are necessary to address associated risks.

Question 2: *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.*

As posted in the meeting 24 hour Summary, available at: https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-20-21-2024-digital-health-advisory-committee-meeting-announcement-11202024#event-materials

The Committee described how pre-GenAI devices have been largely deterministic, focused on retrieval and analysis, while devices enabled with GenAI are probabilistic and generative, and that these relatively unique characteristics should inform risk management of these devices, including employed controls, such as clinical validation and ongoing monitoring. The Committee also stated that GenAI can provide new ways of presenting information that may seem more human-like and give the impression of human intelligence to users. The Committee highlighted how this could contribute to overreliance on the device. The Committee discussed how digital health literacy is essential to consider for patients, particularly in consideration of health equity. Additionally, the Committee noted that the risk of patient harm is a central consideration for risk management and governance. For example, governance may need to differ across different types and implementations of generative AI to leverage site-specific or global governance. The Committee emphasized that clinician training on the use of GenAI-enabled devices is important to support device safety and effectiveness, but not necessarily sufficient to ensure it. The Committee communicated that risk management of these devices should be focused on the risk of patient harm, and they generally agreed on the need for frameworks related to risk management of GenAI-enabled devices, including those focused on the infrastructure needed for deployment in specific settings. The Committee considered benchmarking as a means of comparing capabilities and performance. The Committee suggested that FDA may consider expanding current infrastructure used for other product areas, e.g., those for drugs, to GenAI-enabled devices and posited that new strategies, controls, governance, and frameworks may be needed to mitigate specific GenAI-related risks including those related to ethics and usability. The Committee reiterated how maintaining a human-in-the-loop, with sufficient training, is essential to ensure safety. The Committee discussed the importance of standalone performance testing as well as site-specific clinical validation and ongoing monitoring. Further, the Committee proposed that evaluation of GenAI-enabled devices should include a focus on real world transparency and explainability to the extent possible, as well as real world performance, including the potential variability of that performance in different environments. The committee discussed how plans for monitoring real-world performance should be evaluated at the premarket stage to the extent possible. The Committee highlighted the importance of a shared responsibility for GenAI-enabled devices including manufacturers, health systems deploying the technologies, and health care professionals using them.

Closing Remarks and Adjournment

Mutual gratitude was expressed for everybody's contributions. The November 20 session of the Digital Health Advisory Committee was then officially adjourned.

Contact Information:
Artair S. Mallett
Management Analyst
Center for Devices Radiological and Health
Office of Management
U.S. Food and Drug Administration
Tel: 301-796-9638
Mobile: (301) 538-4714
Artair.Mallett@fda.hhs.gov

I approve the minutes of the meeting as recorded in this summary.

__Ami B Bhatt MD signed electronically 2-5-2025_
Ami Bhatt, M.D.
Chairperson

I certify that I attended this meeting on November 20, 2024
and that these minutes accurately
reflect what transpired.

James P. Swink -S
Digitally signed by James P. Swink -S
Date: 2025.02.13 10:42:03 -05'00'

James P. Swink
Designated Federal Officer