FDA U.S. FOOD & DRUG ADMINISTRATION

24 Hour Summary of the Digital Health Advisory Committee November 20-21, 2024

Introduction:

The Digital Health Advisory Committee to the Food and Drug Administration (FDA) met on November 20-21, 2024, to discuss and provide advice on “Total Product Lifecycle Considerations for Generative AI-Enabled Devices.” Generative AI (GenAI) is defined as the class of AI models that mimic the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content (Source: FDA Digital Health and Artificial Intelligence Glossary – Educational Resource). For the purposes of this Committee Meeting, FDA is using the term “GenAI-enabled device” to refer to a device, as that term is defined in section 201(h) of the Federal Food, Drug, and Cosmetic Act, in which GenAI methods or models are integral to the device’s output or functionality. In other words, for GenAI-enabled devices, GenAI methods or models play a critical role in the device’s primary functions or directly enable its output.

FDA has long promoted a total product life cycle (TPLC) approach to the oversight of medical devices, including artificial intelligence (AI)-enabled devices, and it is committed to advancing regulatory approaches for these devices using current authorities as well as exploring options that may require new authorities. A TPLC approach has become increasingly critical for modern medical devices incorporating technologies that are more complex and intended to iterate faster and more frequently over a device’s life of use than ever before. GenAI-enabled products can be intended to generate new content and provide variable and potentially unbounded outputs for a set of multimodal inputs. Moreover, they frequently rely on complex foundation models that may not themselves be medical devices and may be intended to change rapidly over time. Thus, a TPLC approach is likely important for the management of future, safe and effective GenAI-enabled medical devices.

The questions posed during this meeting aimed to understand the critical information and practices needed for a comprehensive approach to the evaluation and management of GenAI-enabled devices throughout the TPLC. The Committee provided advice and recommendations to FDA on regulatory issues related to GenAI-enabled devices in order to ensure their safety and effectiveness. The meeting focused on three areas: premarket performance evaluation, risk management, and postmarket performance evaluation. Notably, given the variety of distinct intended uses for the many different potential types of GenAI-enabled devices, all the considerations discussed may not be applicable to all GenAI-enabled devices.

FDA Questions and Committee Discussion:

During this two-day meeting, the Committee heard presentations from FDA, industry, academia, health care professionals, patients, and other interested parties.

As noted above, discussion questions on a TPLC approach to GenAI-enabled devices were posed to the committee with a focus on three areas: premarket performance evaluation, risk management, and postmarket performance evaluation.

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.**

   a. **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 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?**

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.

   b. **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?**

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.

c. 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 health care professionals, patients, and caregivers is needed to be conveyed to help improve transparency and/or control these risks?

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.

d. Are there prospective performance metrics that are particularly suited/most informative for these technologies, given their complexity? What kind 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 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.

2. *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.**

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.

3. *Post Market Performance Monitoring:* **Postmarket performance monitoring and evaluation may be important for these devices, particularly because they are non-deterministic. Additionally, after deployment, many generative AI-enabled devices will undergo continuous adjustment based on localized live data, user interactions, and changing conditions. Please discuss the aspects of post market monitoring and evaluation that will be critical to maintaining the safety and effectiveness of these devices.**

    a. **What specific monitoring capabilities should be considered to effectively evaluate and monitor the post market performance of generative AI-enabled devices to ensure they maintain adequate accuracy, relevance, and reliability, especially when adapting to new data?**

The Committee emphasized the importance of postmarket performance monitoring of Gen AI-enabled devices, including automated and scalable approaches to capture how the product is being used, data drift, detection of hallucinations and resulting adverse events. The Committee discussed methodologies including the opportunity for an interim deployment phase before large-scale deployment. The Committee also felt there is a need for specific monitoring capabilities to monitor the postmarket effectiveness of the human-AI interactions. The Committee recognized current FDA postmarket surveillance models and change management approaches as a baseline and encouraged building on existing resources such as Predetermined Change Control Plans (PCCPs) as a useful framework for monitoring devices locally and across multiple sites. The Committee encouraged real world evidence trials to support FDA’s goals to enhance monitoring and evaluation. The Committee also proposed the use of synthetic data as a potential tool for performance evaluation, especially in scenarios with limited data availability. The Committee emphasized the importance of establishing and leveraging standards specific to GenAI, as well as centralized information sharing infrastructure to facilitate reporting back to the manufacturer and broadly to the ecosystem. This may include providing tools to support transparency such as automated user feedback and watermarking to clearly identify the use of GenAI. The Committee discussed including a broad spectrum of reporting considerations, such as error reporting, and implementation failure, as well as clinical outcomes. The Committee called for the need to educate users on postmarket concepts including data drift and to provide clear definition of failure modes and mitigation strategies.

    b. **What specific strategies and tools can be implemented to monitor and manage the performance and accuracy of a generative AI-enabled device implemented across multiple sites, ensuring consistency, and addressing potential regional biases and data variations compared to the device that was authorized?**

The Committee recommended comparing distributions of local data against the training dataset to identify drifts over time. The Committee emphasized the need for automated auditing processes capable of measuring data drift, assessing errors, and identifying corrective actions. The committee was supportive of using approaches such as ensemble methods combining outputs from multiple models to improve robustness, and quality assurance checks embedded within the device. They discussed long-term monitoring of patient outcomes and tracking changes in clinical practices over time. The Committee discussed monitoring for health care professional’s correction of errors in device output.

c. **What methods and metrics can be utilized to effectively monitor and evaluate the post market performance of generative AI-enabled devices that use a multi-layer application design, i.e., the device queries external consumer-grade AI services that are not themselves medical devices?**

The Committee acknowledged the challenge of evaluating GenAI devices that utilize foundation models for which limited information may be available. The Committee encouraged providing information to understand what is in the foundation model, and where that is not possible, noted that it is important to find ways to mitigate the resulting uncertainty. The Committee proposed the need for new tools and frameworks to evaluate and characterize foundation models and their impact or use on a GenAI-enabled device. They emphasized the importance of assessing the representativeness of the training data of the foundation model. The Committee proposed that manufacturers should clearly define performance metrics for each subgroup of the intended use population of the device and monitor outcomes across demographics. The Committee stressed the need for metrics tailored to specific use cases. Beyond model development, the Committee emphasized the need for mechanisms to bring performance data and insights back to community hospitals and medical centers, such as through registries or nonprofit frameworks.

Contact: James Paul Swink
Designated Federal Officer
301-796-6313
James.Swink@fda.hhs.gov

Transcripts, when available, may be downloaded from: https://www.fda.gov/advisory-Committees/advisory-Committee-calendar/november-20-21-2024-digital-health-advisory-Committee-meeting-announcement-11202024#event-materials

OR

Food and Drug Administration
Freedom of Information Staff (FOI)
5600 Fishers Lane, HFI-35
Rockville, MD 20851
(301) 827-6500 (voice), (301) 443-1726