American College of Radiology™

January 21, 2025

Submitted via Regulations.gov

Dockets Management Staff (HFA-305)
Food and Drug Administration
5630 Fishers Lane, Rm. 1061
Rockville, MD 20852

Re: (Docket No. FDA-2024-N-3924) Digital Health Advisory Committee; Notice of Meeting; Establishment of a Public Docket; Request for Comments—Comments of the American College of Radiology

The American College of Radiology (ACR)—a professional association representing more than 40,000 diagnostic radiologists, interventional radiologists, radiation oncologists, nuclear medicine physicians, and medical physicists—appreciates the opportunity to provide written comments to the U.S. Food and Drug Administration (FDA) and its Digital Health Advisory Committee (DHAC) regarding "Total Product Lifecycle Considerations (TPLC) for Generative Artificial Intelligence (GenAI)-Enabled Medical Devices" (FDA-2024-N-3924). The ACR supports the advancement of safe, effective, and clinically useful AI for radiologists to improve care of their patients.

The ACR Data Science Institute (DSI)[1] participated in the Nov. 20-21, 2024, FDA DHAC meeting on GenAI-related TPLC considerations. Attached for the public record are slides presented to DHAC on Nov. 20 by ACR's representative, Bernardo Bizzo, MD, Associate Chief Science Officer, ACR DSI.

General Comments
Radiology devices under 21 CFR Part 892 represent over 75 percent of approximately 1,000 AI-enabled medical devices currently FDA-authorized for the U.S. market, although none to date are specifically GenAI. Most commercially available applications of GenAI are lower risk "non-device" functions that fall outside FDA's jurisdiction (e.g., patient scheduling chatbots used by provider facilities). Nonetheless, it seems likely many clinical functions of GenAI that meet the "medical device" definition (e.g., the use of input image data to generate preliminary/draft report content for radiologist review) will be radiology devices.

The ACR, which established its DSI initiative nearly a decade ago, has expertise in AI implementation and monitoring in radiology provider settings. That experience with current AI-enabled medical devices can help inform the development of effective oversight frameworks for future GenAI in healthcare. We recommend that FDA leverage existing quality assurance (QA) infrastructure, such as the ACR's registries and monitoring programs, in developing its regulatory approaches to GenAI-enabled medical devices.

[1] https://www.acrdsi.org/

As a general approach, the FDA should consider publishing draft guidance on foundation model evaluation and monitoring. This guidance should discuss risk-informed tiers of foundation model oversight. Base foundation models will need appropriate documentation and transparency, and intended use-specific considerations will require performance testing with previously discussed risk-based implementation site validation and monitoring. There should be a standard FDA framework for clinical validation that includes minimum requirements for training data diversity, standardized testing protocols across different clinical scenarios, and performance benchmarks for specific clinical tasks. FDA should also consider healthcare equity-related requirements, including mandatory testing across diverse patient populations, performance monitoring in different healthcare settings, and other bias identification/mitigation approaches.

Specific Comments on Topic 1: Premarket Performance Evaluation

The traditional premarket evaluation approach needs significant adaptation for GenAI-enabled medical devices due to the novel regulatory oversight challenges of this type of technology. Although this distinction is critical for identifying which functions are within or outside FDA's authority, there may be difficulty/ambiguity distinguishing "non-device" GenAI functions from "medical device" functions of the same models.

Device descriptions should include mandatory detailed documentation of foundation model origins and training data to facilitate transparency, fairness, and bias identification and mitigation. The descriptions should specify and delineate between foundation model capabilities and intended use(s). Documentation should also include known limitations and potential failure modes.

Previous experiences with premarket evaluations of radiology devices under 21 CFR Part 892 have shown that studies demonstrating improvement in end-user performance with AI (compared to without AI) are not as useful for demonstrating safety and effectiveness of a device due to the variability and ambiguity of the intended end-user population. These studies can disincentivize innovation, driving industry energy and resources toward functions that do not require these regulatory hurdles as special controls.

Instead, standalone performance testing combined with an explicit description of the intended end-user population (i.e., specialty, board certification, and/or training and experience) should be sufficient for premarket evaluation when combined with post-market local validation and monitoring. The post-market approach should include both traditional metrics for specific intended uses/clinical tasks (e.g., sensitivity/specificity) and new approaches for assessing GenAI outputs. Those new approaches should include comparisons of AI-generated content against expert-validated reference standards, assessments of the hallucination rates in real-world clinical contexts, and evaluations of consistency across similar cases (to address the stochastic outputs by the foundation models).

Specific Comments on Topic 2: Risk Management

Risk management approaches/controls should be stratified based on clinical impact of the device, the qualifications of the device's identified user population, and the ability of a qualified end-user to intervene and mitigate risk before the AI-generated outputs impact patient care decisions. For example, if a GenAI-enabled device function reviews diagnostic image data and generates a preliminary/draft radiology report, the risk of that device can be partially mitigated if a diagnostic radiologist reviews the input imaging data and can modify/correct and approve all AI-generated

content. The risk of the same device is exponentially higher if the AI-generated report is sent to the patient and referring physician directly without qualified end-user review and approval, as most patients and referring physicians do not have sufficient training and experience to identify and mitigate the risks of errors/hallucinations, missed findings, or mischaracterizations of disease.

When stratifying risk, the lowest tier GenAI functions should include lower safety and/or effectiveness concern AI-assisted workflow tools. FDA's medium risk tier should include GenAI-created preliminary/draft content requiring qualified end-user review (i.e., a physician-expert with the clinical qualifications to review and correct/modify the content prior to it influencing patient care). The highest risk tier should include GenAI outputs that directly influence patient care autonomously or semi-autonomously (i.e., with minimal, nominal, or no qualified end-user review).

Specific Comments on Topic 3: Post Market Performance (Evaluation and) Monitoring
GenAI-enabled medical devices introduce novel oversight challenges that require enhanced post-market oversight strategies as part of the TPLC approach. Although expanded statutory authorities may be necessary, FDA should explore implementation specific considerations to ensure safety and effectiveness. These implementation considerations should include site-specific validation requirements to define monitoring cadence, performance comparisons across different sites, and integration with existing QA programs at those sites.

The FDA should explore implementing continuous monitoring requirements that include real-world post-market performance tracking/comparison against premarket benchmarks. Where feasible, comparative evaluations should also be made against nationwide benchmarks leveraging performance data at implementation sites with similar characteristics. This can be enabled by participation in national AI/QA registries, such as the ACR's "Assess-AI" registry[2] discussed during Dr. Bizzo's presentation to DHAC on Nov. 20, 2024. Continuous monitoring should also seek to detect performance drift across different parameters, such as patient population demographics and imaging/input device variations (e.g., manufacturer, version, etc.). Finally, this monitoring should track differences between preliminary/draft content and qualified end-user-approved final content to help identify positive or negative trends in device output accuracy and individual sites/end-users' performance with the device.

The ACR appreciates the time and consideration of FDA staff and DHAC members and welcomes further communication on GenAI and other digital health topics. Please contact Michael Peters, ACR Senior Director, Government Affairs at mpeters@acr.org with questions.

Sincerely,

Dana H. Smetherman, MD, MPH, MBA, FACR
Chief Executive Officer
American College of Radiology

[2] https://www.acrdsi.org/DSI-Services/Assess-AI