Rebooting FDA Regulation of Medical AI

Robert J. Kim*

ABSTRACT

The 1976 Medical Device Amendments established the FDA's authority to evaluate the safety and efficacy of medical devices. While the FDA's initial regulatory framework for devices has remained relatively static over the past five decades, the definition of medical device has expanded dramatically. The same regulatory pathways designed for catheters and leg braces are now being used to assess the safety and efficacy of billion-parameter 'black-box' artificial neural networks. The conceptual connection between regulation and regulated objects has broken, creating a leaky system that fails to account for the unique risks posed by AI systems. A clear indication of this failure is the fact that ninety-six percent of all 'AI-enabled medical devices' have been fast-track approved under a theory of substantial equivalence, a theory that is fundamentally incompatible with the processes and characteristics of machine learning. This article argues that forcing medical AI into the medical device regulatory paradigm is a critical error in both concept and practice, and that continuing to regulate AI as such is a dereliction of the FDA's mandate to ensure the safety and efficacy of medical products. In seeking to remedy this error, this article defines the conceptual boundaries between AI and medical devices, identifies current failure points in medical device regulation, and explores potential directions reforms could take to better execute the FDA's mandate.

DOI: https://doi.org/10.15779/Z384Q7QS6W

* Postdoctoral Fellow in Law and Computer Science, The Ohio State University Moritz College of Law. University of Maryland Francis King Carey School of Law, JD. Vanderbilt University, MS. Brown University BS. I would like to express immense gratitude to Bryan Choi for his mentorship throughout my postdoctoral studies. His guidance and feedback throughout the process of researching and writing of this piece was invaluable. I would also like to thank Ric Simmons for his support throughout my time at Moritz. I would also like to acknowledge Chaz Arnett and Diane Hoffmann, for inspiring me to pursue the study of AI and Health Law. This work was supported in part by the U.S. National Science Foundation under grant CCF-2131531. Any opinions, findings, recommendations, and conclusions expressed in this article are those of the author and do not necessarily reflect the views of the grant sponsor.

TABLE OF CONTENTS

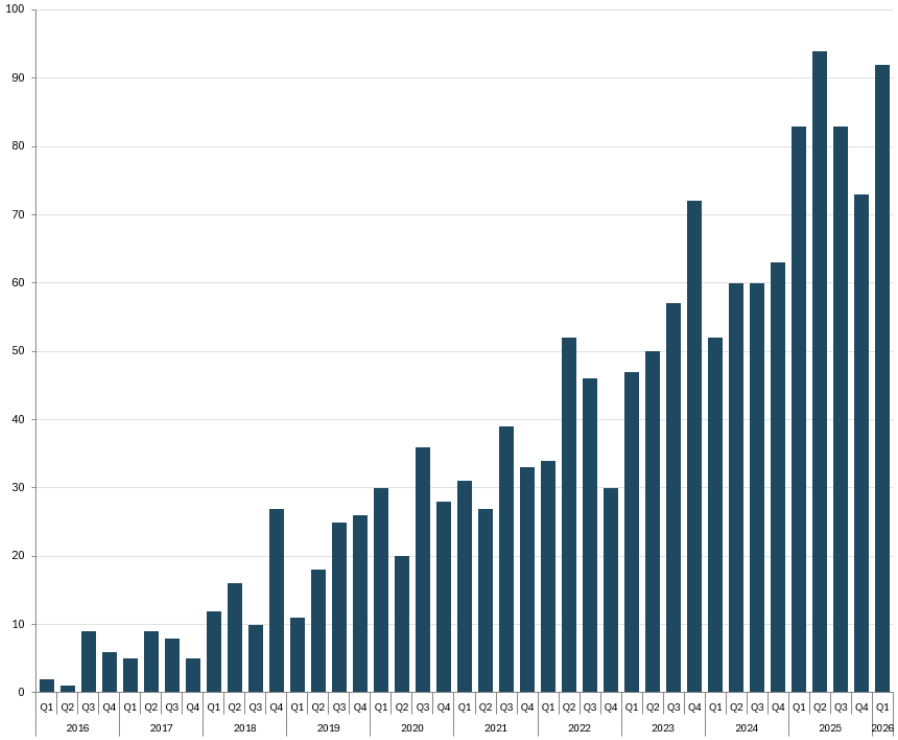
I.
INTRODUCTION

Since the passage of the 1976 Medical Device Amendments (MDA), we have entrusted the Food and Drug Administration (FDA) with ensuring that medical devices entering the market are both safe and effective. Since the inclusion of Medical Artificial Intelligence technologies under the umbrella of medical devices, that trust is being tested. The FDA has approved over 1,200 AI-Enabled Medical Devices[1] so far, with 96 percent approved through the 510(k) substantial equivalence pathway, a pathway specifically designed to eliminate requirements for safety and efficacy testing. Approval rates for AI devices have increased tenfold in the past decade alone, from just 18 approvals in 2016 to 235 approvals in 2024, while the total overall device approval rates have remained relatively static.[2]

1. "AI-Enabled Medical Devices" is the FDA's 2025 successor term to "AI/ML Medical Devices," the term used in pre-2025 literature referenced in this Article. In this Article, the terms "Medical AI" and "AI System" refer to the same technologies as AI-enabled devices, without the "device" framing.

2. *See PMA Approvals*, U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH, https://www.fda.gov/medical-devices/device-approvals-and-clearances/pma-approvals (last visited July 10, 2024); *510(k) Clearances*, U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH, https://www.fda.gov/medical-devices/device-approvals-and-clearances/510k-clearances (last visited July 10, 2024). The number of quarterly device approvals for the 2016–2025 time period has been relatively stable, at about 950 devices per quarter. Approximately 750 approvals were through 510(k), 200 approvals were through PMA (excluding PMA modification

Figure 1: Quarterly device approvals since 2016 (the first year in which at least one AI/ML device was approved each quarter)[3]



This acceleration in AI-enabled device approvals—overwhelmingly through 510(k) substantial equivalence—raises serious questions: Why does the FDA consider AI-enabled devices "substantially equivalent" to pre-AI technologies? Why is the FDA allowing under-validated AI technologies on the market? The answers to these questions are rooted in the fundamental mismatch between AI technologies and the FDA's medical device regulatory framework.

The statutory definition of "medical device" has remained essentially unchanged since 1976,[4] but the reach of the definition has been extended and distorted in ways the drafters of the MDA could never have anticipated. In the

approvals), and 10 were through De Novo. There is a very slight upward trend in 510(k) approvals (1.9 per quarter on average), and a corresponding downward trend in PMAs (1.7 per quarter on average). Data sourced from the openFDA API: https://open.fda.gov/apis/.

3. *See Artificial Intelligence-Enabled Medical Devices*, U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH [hereinafter *AI Medical Devices*], https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices (last visited July 17, 2026).

4. *See* 21 U.S.C. § 321(h)(1) (containing the only post-1976 modification directly relevant to software: "The term 'device' does not include software functions excluded pursuant to section 360j(o) of this title").

late 1980s, the FDA declared that the term "medical device," hitherto used exclusively to describe physical objects and mechanisms, would be applied to intangible instruction sets (i.e., medical software). This decision was conceptually equivalent to classifying methods for peeling potatoes as kitchen appliances. Thirty years later, now that the medical industry has become inured to the conceptual flattening of algorithms and implements, the FDA has again conflated two disparate concepts by declaring medical AI as medical software, and transitively, as medical devices. The defining feature of AI is its reliance on machine-learned (ML) models, which are statistical and not algorithmic constructs. The FDA's definition of what it calls AI Device Software Functions (AI-DSFs) requires us to accept that relational probability distributions are medical devices. This definitional expansion has been problematic because the MDA's frameworks were expressly conceived for the evaluation of physical objects and mechanisms, not intangible algorithms and probability distributions. The MDA's regulatory framework is fundamentally incompatible with its newest regulated objects.

To its credit, the FDA is very cognizant of the unique regulatory challenges that medical AI presents and has been very active in creating AI-specific regulatory guidance.[5] Much of the FDA's work, however, has proceeded in a questionable direction, taking the inherent difficulty of evaluating AI technologies as a reason to shift from a stance of proactive prevention to reactive correction. Over the course of the last decade, the FDA has gradually adopted a "Total Product Life Cycle" (TPLC) approach to device regulation,[6] shifting the assessment of safety and efficacy from premarket testing to postmarket surveillance. The FDA signaled its intention to apply TPLC to AI-enabled devices in a 2019 document, in which it laid out guiding principles for regulating modifications and updates to AI/ML-based Software as a Medical Device (SaMD).[7] The update problem refers to the incompatibility between the rapid, iterative nature of software updates and bug fixes, and the ostensibly heavy burden of premarket approval for seemingly minor changes to software or AI. The ability of the TPLC framework to ensure postmarket safety and efficacy hinges on robust postmarket surveillance mechanisms and timely interventions,

5. *See, e.g.*, U.S. FOOD & DRUG ADMIN., HEALTH CANADA & U.K. MEDS. & HEALTHCARE PRODS. REGUL. AGENCY, TRANSPARENCY FOR MACHINE LEARNING-ENABLED MEDICAL DEVICES: GUIDING PRINCIPLES (2024) [hereinafter TRANSPARENCY REPORT], https://www.fda.gov/media/179269/download; *AI Medical Devices*, *supra* note 3; *Total Product Life Cycle for Medical Devices*, U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH [hereinafter *Total Product Life Cycle*], https://www.fda.gov/about-fda/cdrh-transparency/total-product-life-cycle-medical-devices (last accessed Sep. 6, 2023).

6. *Total Product Life Cycle*, *supra* note 5.

7. *See* U.S. FOOD & DRUG ADMIN., HEALTH CANADA & U.K. MEDS. & HEALTHCARE PRODS. REGUL. AGENCY, PREDETERMINED CHANGE CONTROL PLANS FOR MACHINE LEARNING-ENABLED MEDICAL DEVICES: GUIDING PRINCIPLES (2023) [hereinafter CHANGE CONTROL PLANS], https://www.fda.gov/media/173206/download.

neither of which the FDA has a particularly stellar track record on.[8] The FDA has been explicit in its intention to weaken premarket review in its adoption of TPLC, writing in a 2015 guidance: "[A]pplying postmarket controls to reduce premarket data collection, when appropriate, improves patient access to safe and effective medical devices that are important to the public health."[9]

With sufficient political will and institutional resources, it may be possible for the FDA to better ensure AI-enabled device safety and efficacy via TPLC. The FDA's stated intention to reduce the rigor of an already-weak medical device premarket evaluation system, however, does not inspire trust in the FDA's motives. If the FDA is serious about AI safety, it needs to first acknowledge the fundamental conceptual and regulatory incompatibilities between medical devices and AI systems. The FDA should then consider reclassifying AI systems into a regulatory domain compatible with AI characteristics. Medical AI could potentially be regulated by the Center for Drug Evaluation and Research (CDER) as a drug, or by a sui generis center, possibly one specifically created to address the unique safety and efficacy characteristics of AI systems, especially as AI technologies continue to rapidly evolve.

This Article proceeds in three parts. Part II explains the mismatch between medical AI/ML and traditional devices. Part III unpacks major problems of current FDA regulation of medical devices, including risk assessment, 510(k) approval, and off-label uses. Part IV explores possible paths forward for better medical AI regulation, from incremental reforms to the existing medical device regulatory framework, to alternative frameworks better suited for ensuring AI safety and efficacy.

II.
CATEGORY ERROR: AI IS NOT A MEDICAL DEVICE

A. The FDA's First Misstep: Categorizing Software as a Medical Device

The history of AI's (mis)categorization under medical devices began with the 1976 Medical Device Amendments (MDA) to the Food, Drug, and Cosmetic Act, the foundational legislation that originally established the FDA's mandate over device safety and efficacy. The MDA passed at the beginning of the personal computing revolution, just a year before the release of the Apple II personal computer.[10] The Apple II, Commodore PET, and TRS-80 personal

8. Daniel G. Aaron, *The Fall of FDA Review*, 22 YALE J. HEALTH POL'Y L. & ETHICS 95 (2023); Curt D. Furberg, Arthur A. Levin, Peter A. Gross, Robyn S. Shapiro & Brian L. Strom, *The FDA and Drug Safety: A Proposal for Sweeping Changes*, 166 ARCH. INTERN. MED. 1938 (2006).

9. U.S. FOOD & DRUG ADMIN., CTR. FOR DEVICES & RADIOLOGICAL HEALTH & CTR. FOR BIOLOGICS EVALUATION & RSCH., BALANCING PREMARKET AND POSTMARKET DATA COLLECTION FOR DEVICES SUBJECT TO PREMARKET APPROVAL 6 (2015), https://www.fda.gov/media/88381/download.

10. Pub. L. No. 94-295, 90 Stat. 539 (1976). The Apple II personal computer was released in June 1977.

computers revolutionized the public accessibility and popularity of computing technologies[11] in ways that the drafters of the MDA could not have predicted. Medical software technologies were rapidly developed and deployed, with regulators uncertain how to classify and deal with this new technology.[12]

The FDA ultimately chose to regulate medical software as medical devices, despite clear and significant conceptual, definitional, and practical incompatibilities. By the mid-1980s, software had significant penetration in medical practice. Regulators at the FDA knew software was different from traditional devices but, given the functional relationship between software and medical hardware, ultimately chose to conceptually equate the software with hardware (devices).[13] This decision was not without its critics; in testimony to the House Committee on Science and Technology in 1986, Vincent Brannigan[14] stated of FDA personnel and their treatment of medical software:

> I think that their idea is that this is some kind of new bedpan, that it's a thing, and their attitude is oriented toward it as a thing . . . they're wrestling with the fact that it's very hard to mesh this very strange thing in with the rest of their other things. So, rather than try to create a separate regulatory program, they just say, "Well, we're going to treat it the same as all the other things."[15]

This statement, while perhaps unfair in its characterization of FDA personnel, was not an inaccurate description of the FDA's institutional treatment of software-based medical products. The FDA released its first "Draft FDA Policy for the Regulation of Computer Products" in 1987 and updated it in 1989.[16] The guidance was the first official statement of intention to classify some types of standalone medical software as medical devices, further classifying software "components" of medical devices as medical devices. Responding to the release of this guidance, Brannigan argued that while software contained in medical devices was "clearly" a medical device, the statutory definition of

11. Jeremy Reimer, *Total Share: 30 Years of Personal Computer Market Share Figures*, ARS TECHNICA (Dec. 14, 2005), https://arstechnica.com/features/2005/12/total-share/.

12. *See generally* Vincent Brannigan, *The Regulation of Medical Expert Computer Software as a "Device" Under the Food, Drug, and Cosmetic Act*, 27 JURIMETRICS 370 (1987); Nathan Cortez, *Regulating Disruptive Innovation*, 29 BERKELEY TECH. L.J. 175 (2014); Nathan Cortez, *Digital Health and Regulatory Experimentation at the FDA*, 21 YALE J.L. & TECH. 4 (2019).

13. Cortez, *Digital Health*, *supra* note 12.

14. Brannigan was, at the time, a professor in the Department of Textiles and Consumer Economics at the University of Maryland. Brannigan, *supra* note 12, at 370.

15. *Information Technologies in the Health Care System: Hearing Before the Subcomm. on Investigations & Oversight of the H. Comm. on Sci. & Tech.*, 99th Cong. 167 (1986) [hereinafter *Hearing*] (testimony of Vincent Brannigan, Assoc. Professor, Dep't of Consumer Econs., Univ. of Md.), https://files.eric.ed.gov/fulltext/ED281423.pdf.

16. U.S. Food & Drug Admin., Draft Policy Guidance for Regulation of Computer Products; Availability, 52 Fed. Reg. 36104 (Sep. 25, 1987); U.S. Food & Drug Admin., Draft Policy for the Regulation of Computer Products (Nov. 13, 1989).

medical device was inapposite to software that uses "stored algorithms and data to evaluate patients and treatment, but does not directly examine or treat the patient." [17] Brannigan focused on, among other attributes, the physical intangibility of software as separating it from every other type of medical device: "Software is intangible, somewhere between an idea and the expression of an idea."[18]

Although Brannigan's contentions seem sensible in retrospect, the ontological distinction between software and physical hardware was not a settled issue in the 1970s and 1980s. In 1978, philosopher James Moor described a computer program as "a set of instructions which a computer can follow (or at least there is an acknowledged effective procedure for putting them into a form which the computer can follow) to perform an activity." [19] Moor did not, however, explicitly equate "computer programs" with "software," arguing that the definitions of "software" and "hardware" overlap depending on user and usage context.[20] His framing emphasized the physical embodiment of software over its informational nature, a framing colored by some contemporary computers that were still, in Moor's words, programmed by "plugging in wires and throwing switches."[21] In 1988, Peter Suber took a more bird's-eye-view conceptual framing, choosing to define software as a "*pattern* that may be represented in many different materials."[22] By Suber's definition, there was no conceptual difference between software and hardware.

Decades later, in 2017, having the benefit of retrospect on multiple decades of hardware and software co-evolution, William Duncan proposed definitions of software and hardware that are more aligned with current practical realities of computing technologies. Duncan criticized Suber's definition as ridiculously overbroad, opining that the idea of a peanut butter sandwich would count as

17. Brannigan, *supra* note 12, at 371–72; *Hearing*, *supra* note 15, at 83.

18. Brannigan, *supra* note 12, at 371–74. Brannigan refers to these software as "medical expert systems." Brannigan also references the unpublished opinion in *USA v. Undetermined Quantity of Articles of Device*, which determined that items that only transfer information were not within the statutory definition of "device." No. G80-926 (W.D. Mich. Nov. 23, 1982).

19. James H. Moor, *Three Myths of Computer Science*, 29 BRIT. J. PHIL. SCI. 213, 214 (1978). Moor anticipates the applicability of this definition to AI systems: "In principle virtually any physical entity or process could be understood symbolically as a computer instruction if one imagines an appropriate computer which could understand and execute the instruction. But as a practical matter what we regard as computer instructions, and consequently what we regard as computer programs, is determined by the computers available." *Id.* at 215.

20. *Id.* at 215. Moor defines software "[f]or a given person and computer system" as something that "will be those programs which can be run on the computer system and which contain instructions the person can change." *Id.* Conversely, "the hardware will be that part of the computer system which is not software." *Id.* Paraphrasing his argument: As one moves along the "user" spectrum from computer manufacturer, to programmer, to consumer, more and more of the computer system (hardware plus software) becomes "hardware" as the scope of user's ability to change the system decreases.

21. *Id.*

22. Peter Suber, *What Is Software?*, 2 J. SPECULATIVE PHIL. 89, 104 (1988) (emphasis added). Suber further contends that neural patterns are software but have so far not been "liftable" (i.e., copiable) due to the extreme complexity of the pattern. *Id.*

software under Suber's "pattern"-based definition.[23] Duncan departed from both Moor and Suber, recognizing software as *instructions* embedded in a machine-readable *artifact*, and hardware as *artifacts* that *bring about results* of calculations.[24] This definition places primacy on software's instructional characteristics as separating it from hardware's functional operations. Duncan's modern definitions align well with Brannigan's contentions thirty years prior—a post-hoc reinforcement of the credibility of Brannigan's framings. Still, the FDA was ultimately unpersuaded, whether by ontological arguments or the draw of regulatory convenience. The FDA chose to regulate medical software as medical devices, based on the existence of a functional relationship between software and hardware.[25]

Even if the FDA genuinely believed in the equivalence of software and hardware, it should have recognized that even basic medical software adds significant complexity to any system it is part of, far beyond the complexity of traditional devices. The rapid mutability of software adds a compounding concern, in that seemingly benign modifications to software can create unpredictable cascading errors. The early-1980s fatalities caused by the infamous Therac-25, a software-controlled radiation therapy device, were directly attributed to naïve porting of software from a previous device without fully evaluating possible incompatibilities. This series of fatalities should have at the very least spurred FDA efforts to craft more rigorous software testing standards for devices with the capacity of inflicting such grievous harm, but subsequent nonbinding software guidance over the next few decades did little to practically address software complexity.[26]

Decades after the FDA made its choice to regulate medical software as devices, the term "Software as a Medical Device" (SaMD) was semi-formally defined in a guidance document released by an International Medical Device Regulators Forum (IMDRF)[27] working group chaired by the FDA.[28] The

23. William D. Duncan, *Ontological Distinctions Between Hardware and Software*, 12 APPLIED ONTOLOGY 5, 13 (2017); Nurbay Irmak, *Software Is an Abstract Artifact*, 86 GRAZER PHILOS. STUD. 55 (2012).

24. Duncan, *supra* note 23, at 5 ("A software program is a specification that consists of one or more programming language instructions and whose concretization is embodied by an artifact that is designed so that a physical machine may read the concretized instructions, whereas hardware is an artifact whose functions are realized in processes that directly or indirectly bring about the result of some calculation"). Duncan draws the notions of "specification" and "concretization" from Raymond Turner and Timothy R. Colburn respectively. *See generally* Raymond Turner, *Specification*, 21 MINDS & MACHS. 135 (2011); Timothy R. Colburn, *Software, Abstraction, and Ontology*, 82 MONIST 3 (1999).

25. E. Stewart Crumpler & Harvey Rudolph, *FDA Software Policy and Regulation of Medical Device Software*, 52 FOOD & DRUG L.J. 511 (1997) (explaining that the FDA's "device" classification was based on software's functional relationship to physical medical devices and an interpretation of the term "contrivance").

26. Cortez, *Regulating Disruptive Innovation*, *supra* note 12.

27. The International Medical Device Regulators Forum is a consortium of international regulatory authorities, of which the FDA is a member.

28. As international regulatory guidance, IMDRF documents are not binding on the FDA.

definition explicitly included standalone software[29] within the proposed definition of SaMD: "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device."[30] The IMDRF also established a working definition of "medical device" to anchor the definition of SaMD. This definition was nearly identical to the U.S. Code definition of medical device, save for a couple of significant edits: the IMDRF (1) explicitly included the term "software" as a category of device;[31] and (2) removed the term "contrivance," a term that has been salient to arguments about software's status.[32] These targeted divergences from the otherwise nearly-verbatim copy of the U.S. Code's definition of a medical device suggest that the IMDRF considered the U.S. Code's definition inadequate to cover software.

Beyond the conceptual, practical, and definitional issues with regulating software as medical devices, the FDA's choice also created an unfortunate precedent regarding device relationality: assigning "device" designation based on functional relationships between entities, rather than by common characteristics inherent to these entities. The FDA used software's functional relationship with computer hardware as a primary justification to categorize software as a device.[33] The FDA is now taking this relational justification one conceptual degree of separation further with AI, regulating AI as a device based on (ML) statistical models' functional, rather than characteristic-based, relationships to software algorithms.[34]

B. An Illusory Bridge: Software Does Not Join Devices with AI

Justifiable or not, software now has a firm place in the canon of medical devices by dint of time and praxis, if not by conceptual justifiability. In the absence of specific statutory directives, the FDA is treating AI-enabled medical systems as medical devices[35] based on functional relationships between AI

29. I.e., intangible software, which is the type of software that Brannigan argued against inclusion of.

30. IMDRF SaMD Working Group, *Software as a Medical Device (SaMD): Key Definitions*, INT'L MED. DEVICE REGULS. F. 6 (Dec. 9, 2013), https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf. Notably, the definition of "Medical Device" suggested by this document explicitly adds the term "software" to its definition that was otherwise identical to the U.S. Code's statutory definition, which could be interpreted as a tacit acknowledgement that the U.S. Code definition was not applicable to software.

31. *Id.* ("any instrument, apparatus, implement, machine, appliance, implant, reagent for in vitro use, software").

32. Brannigan, *supra* note 12, at 373, and Crumpler & Rudolph, *supra* note 25, at 511, refer to the term "contrivance" as covering software (with diverging levels of credulity).

33. Crumpler & Rudolph, *supra* note 25, at 511–13.

34. *AI Medical Devices*, *supra* note 3.

35. *See* U.S. FOOD & DRUG ADMIN., CTR. FOR DEVICES & RADIOLOGICAL HEALTH, CTR. FOR BIOLOGICS EVALUATION & RSCH., CTR. FOR DRUG EVALUATION & RSCH., OFF. OF COMBINATION PRODS. IN THE OFF. OF THE COMM'R, MARKETING SUBMISSION RECOMMENDATIONS FOR A PREDETERMINED CHANGE CONTROL PLAN FOR ARTIFICIAL INTELLIGENCE-ENABLED DEVICE

systems and software.[36] Software is not a salient conceptual bridge between medical AI and the physical devices envisioned by the MDA's drafters.

James Moor described an essential conceptual distinction between computer programs and computer models in 1978:

> [A] computer model is usually much more than just a computer program, for a computer program is a set of computer instructions in the sense defined earlier. For the computer model to be understood as a model the isomorphism has to be pointed out between the computer activity (as produced by the program, perhaps) and the subject matter in question. In general, one must give the computer activity a symbolic interpretation which goes beyond the standard symbolic understanding of the program in order to make it a model.[37]

To paraphrase, models contain meaning beyond their relevance to the algorithmic execution of a program; they derive their meaning and usefulness as representations of concepts external to the program.

The FDA seems to agree with these conceptual distinctions, given how it defines AI terms. The agency offers the following definitions of AI concepts:

> **Artificial Intelligence** is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models

SOFTWARE FUNCTIONS (2025) [hereinafter *Recommendations*], https://www.fda.gov/media/166704/download.

36. *See Artificial Intelligence in Software as a Medical Device*, U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH (2025) [hereinafter, *AI in Software*], https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device (The FDA's informational pages on AI/ML devices are subsections of their page on SaMD).

37. Moor, *supra* note 19, at 220. As an interesting sidenote, in the previous paragraph Moor discusses models without underlying theories: "[S]oap film on irregularly shaped wires is sometimes regarded as an analogue computer which gives solutions to minimisation problems (Isenberg [1976]). Although the soap film on the irregularly shaped wire can serve as a model, it does not generate a theory about minimisation. Today there is no mathematical theory which can provide a general existence proof for such solutions let alone an analytical approach for producing them. Such a mathematical theory might be developed some day." *Id.* Ten years later, in 1989, George Cybenko demonstrated such an existence proof now known as the "universal approximation theorem" that proved for a function *f*, there exists a sequence of neural networks that will approximate that function to an arbitrary criterion of closeness. George Cybenko, *Approximation by Superpositions of a Sigmoidal Function*, 2 Math. Control Signals Sys. 303, 303–04 (1989); *see also* Kun Wang, *A Note on Cybenko's Universal Approximation Theorem*, ARXIV 1, 1 (Aug. 26, 2025), https://arxiv.org/abs/2508.18893. In other words, basically every system can be modeled accurately by a neural network. The catch is, practically speaking, we have no idea whether our ML training algorithms can create these ideal networks, nor can we confirm a given network is an ideal solution to a problem.

> through analysis in an automated manner; and use model inference to formulate options for information or action.
>
> **Machine Learning** is a set of techniques that can be used to train AI algorithms to improve performance at a task based on data.
>
> **Machine Learning Model:** A mathematical construct that generates an inference or prediction for input data. This model is the result of an ML algorithm learning from data. Models are trained by algorithms, which are step-by-step procedures used to process data and derive results. AI systems (e.g., AI-enabled medical devices) employ one or more models to achieve their intended purpose.[38]

Although the FDA definition of AI implies a very expansive range of potential capabilities and complexity, it has useful limiting language, the crucial phrases being "machine-based," "abstract such perceptions into models," and "use model inference to formulate options."[39] Also, in defining "Machine Learning Model," the FDA states that AI systems employ one or more Machine Learning models, implying that for the FDA's purposes, an AI system is only AI if it incorporates ML model(s), creating a required dependency of "AI" systems on the use of ML. As for "abstracting perceptions" into models and performing "model inference," the FDA appropriately refers to models as mathematical constructs rather than as algorithms:

> A model is a mathematical construct that generates an inference or prediction based on new input data. In this guidance, when "AI-enabled device" is used, it refers to the whole device, whereas when "AI-DSF" ["AI-Device Software Function"] is used, it refers only to the function that uses AI. In this guidance, when "model" is used, it refers only to the mathematical construct.[40]

The FDA's definition of ML model also creates a sharp distinction between the software processes that create ML models (ML algorithms) from the ML

38. *AI in Software*, *supra* note 36; *FDA Digital Health and Artificial Intelligence Glossary – Educational Resource*, U.S. FOOD & DRUG ADMIN. (last updated Sep. 26, 2024), https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource#m.

39. *AI in Software*, *supra* note 36.

40. Draft Guidance, *Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations*, U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH, CTR. FOR BIOLOGICS EVALUATION & RSCH, CTR. FOR DRUG EVALUATION & RSCH., OFF. OF COMBINATION PRODS. IN THE OFF. OF THE COMM'R 2–3 (2025) [hereinafter *Lifecycle Management*], https://www.fda.gov/media/184856/download.

models themselves. All current medical AI models are "locked" models.[41] This means that once a model is trained, the model's use is entirely independent of its training algorithm, in much the same way that a published book's existence is independent of future edits by its author. Furthermore, ML model architectures are not defined by their training algorithms. Multiple training algorithms exist for any given ML architecture.[42] These algorithms are tools to train ML models, not components of the models themselves.

Regarding the algorithmic nature of generating inferences from ML models, one might argue that because the process of inferencing consists of traversing directed graphs in a series of ordered steps, ML models are effectively software algorithms. While this may be arguable from a naïve structural standpoint, ML models do not derive their core functionality from their inferencing algorithms. Running an inferencing algorithm without the model's learned statistical parameters has no functional value, akin to booting a computer without an operating system.

Perhaps the most critical difference between software and AI is that software is architected mechanistically, variable by variable, function by function, to achieve a particular end. The algorithms used to build software are determined by their intended function(s). In contrast, for the core of AI systems—the ML model(s)—model architecture is not determinative of function.[43] Functions are mapped into ML models through the learning process. Of course, particular model architectures may have properties better suited to learning particular types of *mathematical* functions, but this has very little bearing on the suitability of an architecture for the ultimate purpose of the AI system.

For example, in the field of generative AI, two of the most popular AI architectures are "diffusion" models and "transformer" models. Diffusion

41. Sara Gerke, *Health AI for Good Rather Than Evil? The Need for a New Regulatory Framework for AI-Based Medical Devices*, 20 YALE J. HEALTH POL'Y L. & ETHICS 433, 496–98 (2021) (contrasting locked models from so-called "adaptive" models that continually learn through usage; as of this writing, model and algorithm combinations capable of true continuous learning do not exist and are still open problems in machine learning).

42. *See, e.g.*, Ruslan Salakhutdinov & Geoffrey Hinton, *Deep Boltzmann Machines*, *in* PROC. TWELFTH INT'L CONF. ON A.I. & STAT. 448 (2009), https://proceedings.mlr.press/v5/salakhutdinov09a.html (describing a new algorithm for training Deep Boltzmann Machines).

43. The convolutional neural network architecture has been used for (respectively) speech recognition, classification of neutrino collisions and, its original and most prevalent purpose, image recognition tasks. *See, e.g.*, Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn & Dong Yu, *Convolutional Neural Networks for Speech Recognition*, 22 IEEE/ACM TRANS. AUDIO, SPEECH AND LANG. PROC. 1533 (2014); R. Abbasi, M. Ackermann, J. Adams, S.K. Agarwalla, J.A. Aguilar, M. Ahlers, J.M. Alameddine, N.M. Amin, K. Andeen, G. Anton, C. Argüelles, Y. Ashida, S. Athanasiadou, S.N. Axani, X. Bai, A. Balagopal V., M. Baricevic, S. W. Barwick, V. Basu, R. Bay et al., *Observation of Seven Astrophysical Tau Neutrino Candidates with IceCube*, ARXIV (Mar. 26, 2024), https://arxiv.org/abs/2403.02516; Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard & L.D. Jackel, *Backpropagation Applied to Handwritten Zip Code Recognition*, 1 NEURAL COMPUTATION 541 (1989).

models were originally designed for image generation.[44] Image Diffusion can be described as a "denoising" process, where an image starts as pure noise and is iteratively "denoised" by ML model inferencing until a noise-free image is formed.[45] Transformer models, the underlying architecture of the majority of current large language models (LLMs), use "attention" mechanisms to learn sequential relationships in data. The attention mechanisms relate the relative importance of the positions of items in a sequence to determine the semantic value of a sequence (in encoder models) or to predict the next word in a sequence (decoder models).[46] Preferentially using denoising models for images, while using sequence prediction models for language, makes intuitive sense. This intuition, however, is an artifact of human experience, rather than a reflection of actual ML model architectural capabilities. There are now diffusion LLMs that generate text by non-sequentially "denoising" initially random text sequences into coherent, structured language.[47] Writing an essay word by word in a random, nonlinear order[48] may be alien to human intuition, but perfectly sensible to an ML model trained for such a method. Both Apple and Google are pursuing text diffusion as a potentially more computationally efficient method for text generation than transformers.[49] On the flip side, vision transformers are now used to both interpret and generate images as sequences,[50] often with better alignment between prompt and image than diffusion models. This was popularly demonstrated during the ChatGPT Studio Ghibli trend, following OpenAI's

44. *See* Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser & Björn Ommer, *High-Resolution Image Synthesis with Latent Diffusion Models*, ARXIV (Apr. 13, 2022), https://arxiv.org/abs/2112.10752.

45. Jonathan Ho, Ajay Jain & Pieter Abbeel, *Denoising Diffusion Probabilistic Models*, *in* 33 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 6840 (2020), https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

46. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin, *Attention Is All You Need*, *in* 30 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (2017), https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

47. *See, e.g.*, Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li & Lingpeng Kong, *Dream 7B,* HKU NLP GROUP, https://hkunlp.github.io/blog/2025/dream/ (last visited Aug. 6, 2025); *DiffusionGemma*, GOOGLE DEEPMIND, https://deepmind.google/models/gemma/diffusiongemma/ (last visited July 17, 2026).

48. "Random order" means, for example, first writing word 56, then word 2, then word 128, until all the words are filled in.

49. *See Gemini Diffusion*, GOOGLE DEEPMIND, https://deepmind.google/models/gemini-diffusion/ (last visited Aug. 6, 2025); Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong & Yizhe Zhang, *DiffuCoder: Understanding and Improving Masked Diffusion Models for Code Generation*, ARXIV 1, 1, 4 (June 26, 2025), https://arxiv.org/abs/2506.20639. (Apple's model is based on Huawei's Dream 7B).

50. Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu & Chong Ruan, *JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation*, DEEPSEEK (Mar. 24, 2025), https://arxiv.org/abs/2411.07975.

reveal of its vision transformer implementation.[51] The human equivalent of sequence-based image creation would be drawing an image pixel by pixel, left to right, row by row.

Hopefully, these examples demonstrate that an ML model's architecture provides almost no information regarding its function. Once trained, a model's function can be evaluated, but not mechanistically (discussed further in the next Section). Evaluation of the safety and efficacy of an AI system cannot be accomplished by the same methods used for software.

Even though the FDA's definitions show that the FDA clearly recognizes the definitional and conceptual difference between software and AI, the FDA has chosen to ignore these differences, bridging software with AI through the concept of an AI-Device Software Function (AI-DSF).[52] Casting ML models as "Software Functions" subordinates the importance of the ML model to the software built around it. This is entirely backwards with respect to the relative importances of the ML model and the software that interfaces with it. Regulation should target the aspects most salient to the behavior of a product, not a superficial wrapper. Regulating AI devices by the metrics applied to software functions is conceptually equivalent to regulating drugs by assessing the safety of the capsules that contain medication, rather than by the safety of the medication itself.

At a very broad level of abstraction, ML models could conceivably be described as software functions—artifacts that convert inputs into outputs. But the functional aspect of a method of medical treatment is only relevant to the purpose of the treatment, not to its safety or effectiveness. Software code can be reviewed deterministically to assess whether it processes salient inputs into proper outputs. For ML models, there is no simple way to determine if models are even considering critical inputs at all.[53]

C. A Meaningful Boundary Between Software and AI

Given the clear differences between SaMD and AI systems, drawing a regulatory line between the two seems relatively straightforward. There are clear conceptual differences between software and ML models: software embodies algorithmic/instructional information, while ML embodies statistical information. These conceptual differences, however, only matter to the degree that they may necessitate different types of evaluations for safety and efficacy.

The primary concern generally attributed to medical AI is the problem of "black-box medicine," a term coined by W. Nicholson Price II to describe

51. *See* Dani Di Placido, *The AI-Generated Studio Ghibli Trend, Explained*, FORBES (Mar. 27, 2025), https://www.forbes.com/sites/danidiplacido/2025/03/27/the-ai-generated-studio-ghibli-trend-explained/.

52. *See Lifecycle Management*, *supra* note 40.

53. Alex J. DeGrave, Joseph D. Janizek & Su-In Lee, *AI for Radiographic COVID-19 Detection Selects Shortcuts Over Signal*, 3 NAT. MACH. INTEL. 610 (2021).

medical technologies that make determinations in ways that are not human-understandable, even to the creators of these technologies.[54] To be clear, the lack of understanding of AI is not one of mechanistic complexity; the mathematical principles behind how and why ML models work are well understood. Nor is interpretability solely a function of input dimensionality; traditional software processing often involves high-dimensional input.

The key factor that makes AI models black boxes is the concept of nonlinearity. Most well-known methods of modeling data use linear statistical techniques. For example, linear regression (e.g., line of best fit) could technically be described as a machine learning technique, since a line of best fit is a function 'learned' from a training set, e.g. a scatter plot. Unlike nonlinear models, however, linear regression produces models with one-to-one relationships from input to output that are immediately decomposable into simple sums of linear functions.[55] Other linear ML models generated through methods like Principal Component Analysis[56] or Linear Support Vector Machines[57] are similarly conducive to component-level interpretation. These methods of ML would likely not be considered black-box because their outputs can be analyzed as individual components. On the other hand, nonlinear statistical models like Multi-Layer Perceptrons[58] do not have one-to-one relationships between inputs and outputs. The layers of neurons in MLPs are fully connected, meaning every neuron in one layer is connected to every neuron in the layer(s) before and after it. This means that the individual contribution of a neuron in a layer becomes inseparably entangled with all its neighbors' contributions just two layers deep.

Another practical separation between linear and nonlinear models is the predictability of training. Due to the commutative property, linear (and polynomial) regressions will always have only one result, no matter what order the training data is presented. In contrast, nonlinear models are very sensitive to the order of data presentation.[59] Given a training set, linear models will always deterministically produce a single result. Training a nonlinear model can have

54. *See generally*, W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J. L. & TECH. 419 (2015); W. Nicholson Price II, *Describing Black-Box Medicine*, 21 B.U. J. SCI. & TECH. L. 347, 429–31 (2015).

55. *See, e.g.*, Thomas M.H. Hope, *Chapter 4 - Linear Regression*, *in* MACHINE LEARNING 67 (Andrea Mechelli & Sandra Vieira eds., 2020), https://www.sciencedirect.com/science/chapter/edited-volume/abs/pii/B9780128157398000043.

56. *See, e.g.*, Ian Jolliffe, *Principal Component Analysis*, *in* INTERNATIONAL ENCYCLOPEDIA OF STATISTICAL SCIENCE 1094 (2011), https://link.springer.com/rwe/10.1007/978-3-642-04898-2_455.

57. *See, e.g.*, William S. Noble, *What Is a Support Vector Machine?*, 24 NATURE BIOTECH. 1565 (2006).

58. Marius-Constantin Popescu, Valentina E. Balas, Liliana Perescu-Popescu & Nikos Mastorakis, *Multilayer Perceptron and Neural Networks*, 8 WSEAS TRANSACTIONS ON CIRS. & SYS. 579 (2009).

59. *See* Jeremy Mange, *Effect of Training Data Order for Machine Learning*, 2019 INT'L CONF. ON COMPUTATIONAL SCI. AND INTEL. (CSCI) 406 (2019), https://ieeexplore.ieee.org/abstract/document/9071297.

countless results depending on training rate and method, some of which result in useful models, and many which do not.

Finally, the weights of linear models can be analyzed mechanistically, as each linear weight has effects that are independent and separable from those of all the other weights. In nonlinear models, the interdependence of the effects of any given weight on many other weights in the models makes the measurement and interpretation of individualized contributions intractable.

Scholars have proposed other boundaries for separating AI from non-AI systems. For example, Sara Gerke, who has written extensively on AI medical devices, suggests that lines could potentially be drawn at thresholds of interpretability.[60] The general idea is that if an AI system is interpretable, then it should be subject to a lower level of scrutiny than an AI system that is not interpretable. Gerke also suggests that, given equivalent levels of performance, interpretable models should be favored over non-interpretable models. Interpretability is unfortunately a highly context-dependent concept. Some tie the definition to actual human capacity to understand an ML model.[61] Others define interpretability as the ability to provide a human with the illusion of understanding, such as creating a diagnostic factor-based scoring mechanism generated by uninterpretable nonlinear methods.[62]

To effectively distinguish AI systems that present "black-box" concerns from traditional software, we can treat AI devices that utilize traditional linear statistical models as not presenting concerns that go beyond those of traditional SaMD. Mechanistically interrogating the characteristics of a linear model is at least as straightforward as reviewing the code of a SaMD device.[63] On the other hand, medical systems that incorporate nonlinear statistical models are resistant to mechanistic evaluation, requiring empirical evaluation instead.

60. Gerke, *supra* note 41, at 511–12.

61. *See, e.g.*, Dylan Slack, Sorelle A. Friedler, Carlos Scheidegger & Chitradeep Dutta Roy, *Assessing the Local Interpretability of Machine Learning Models*, ARXIV (Aug. 2, 2019), https://arxiv.org/abs/1902.03501.

62. *See, e.g.*, Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet & Tom McGrath, *Open Problems in Mechanistic Interpretability*, ARXIV (Jan. 27, 2025), https://arxiv.org/abs/2501.16496.

63. Most software is mathematically nonlinear in design and execution. The difference between software nonlinearity and ML model nonlinearity, however, is that software is intentionally designed, so the purpose and behavior of all components are known. Nonlinear statistical models are trained by learning algorithms, and the learned parameters are not intentionally designed with known functions.

III.
AI UNCHECKED: THE INADEQUACY OF DEVICE REGULATION

The medical device regulatory system emerged in an era when nearly all medical devices were physical constructions, each designed to fulfill a particular set of mechanical or energetic functions. Assessing risks of such devices was relatively straightforward because form tends to follow function for physical devices. Similarly, when determining 510(k) substantial equivalence, [64] comparisons between devices were straightforward based on clear similarities in design. Medical AI is impossible to analyze in these ways. As discussed, *supra*, the architecture of ML models does not dictate their function, and architectural similarity between AI models does not indicate similarity of behavior or performance. Classic mechanistic analyses simply do not apply to nonlinear statistical systems. Additionally, the "streamlining" of device review processes over time has further reduced the effectiveness of review for medical AI. Systemic structural issues in the FDA's system of device risk classification,[65] the 510(k) substantial equivalence pathway,[66] and the unique risks AI brings to off-label usage have resulted in medical AI devices effectively bypassing systems meant to evaluate and ensure the safety and efficacy of medical devices.

A. FDA Device Risk Classification: A Metric That Lost Its Meaning

The first step device manufacturers engage with in the FDA's approval process is risk classification. Initial risk classification is critical because it is the gatekeeper that determines the level of regulatory scrutiny that a device must undergo post-classification. Originally created as a holistic assessment, risk classification has since evolved into an indication-and-purpose-based assessment that critically fails to consider the relevance of new, unproven technologies like AI to the evaluation of risk.

The 1976 Medical Device Amendments separated devices into three tiers of risk, each with different approval requirements.[67] Class I encompasses low-risk devices requiring only general controls[68] for assurance of safety. Class II refers to medium-risk devices whose safety and efficacy cannot be ensured by general controls, requiring category-specific special controls. Class III represents the highest-risk devices whose risks cannot be adequately addressed by general or special controls, and instead require full premarket clinical testing. The

64. Discussed *infra* Section III.B.

65. Susan Bartlett Foote, *Loop and Loopholes: Hazardous Device Regulation Under the 1976 Medical Device Amendments to the Food, Drug and Cosmetic Act*, 7 ECOLOGY L.Q. 101 (1978).

66. Gerke, *supra* note 41.

67. It would have likely been unreasonable to require full premarket clinical trials of bedpans, patient gowns, and adhesive bandages.

68. e.g., Quality Management System Regulation (QMSR), that do not regulate the design of a product, but rather the methods of manufacture, packing, storage, etc. 21 C.F.R. pt. 820, https://www.federalregister.gov/documents/2024/02/02/2024-01709/medical-devices-quality-system-regulation-amendments#h-47.

intention behind this three-tier risk framework was to streamline approval of low-risk devices and free up developer and regulatory resources for testing higher-risk devices. Effectiveness in practice, however, depends on the common conceptions of risk being aligned with the criteria for measurement of risk.

In the early days of the MDA, risk classification was assessed holistically based on the individual characteristics of individual devices. In the months leading up to the passage of the MDA, a battery of eighteen questions was developed for device classification panels to determine initial device risk classifications.[69] There was little guidance and consistency among the numerous device panels on how to weigh the value of each question, leading to inconsistent application.[70] Still, at the very least, holistic assessments were conducted per device. Devices are no longer evaluated in this way; evaluation of risk is now category-based, determined primarily by two inquiries: a device's "intended use" and its medical "indications for use."[71] Applicant manufacturers determine which (of several thousand) device categories their device likely belongs in.[72] If a candidate device reasonably fits into a category—for example, software designed to process x-rays might fit into the category "Medical Image Analyzer"—the manufacturer proposes that category to the FDA with supporting documentation. If the FDA does not dispute the classification, the device receives the risk level assigned to that device category. Technological characteristics affecting the safety and efficacy of devices are supposed to be considered in classification,[73] but these differences rarely determine product

69. Medical Device Classification Procedures: Notice to Manufacturers, 40 Fed. Reg. 21849–50 (U.S. Food & Drug Admin. May 19, 1975). Here is a subset of the questions that seem particularly relevant to considering current AI/ML systems: "3: Is the device life-sustaining or life-supporting? . . . 4: Is the device or diagnostic information derived from use of the device potentially hazardous to life or good health when properly used? . . . 5: Is the device of such a nature that: (a) sufficient scientific and medical data exist from which adequate standards governing the device safety and efficacy could now be established; and; (b) development and application of such a standard would be adequate to control the device? . . . 15: If the device performs some measurement function, should the accuracy, reproducibility or limitations of the information supplied be clearly indicated to the user by appropriate labeling, instructions, or precautions? . . . 16: Does the device have performance characteristics which should be maintained at a satisfactory level, such level having general agreement among the user groups? . . . 17: Is the device used with other devices in such a way that the system in which it is used can be hazardous if the system is not assembled, used or maintained in a satisfactory fashion?"

70. Foote, *supra* note 65, at 121.

71. Crumpler & Rudolph, *supra* note 25, at 511 (mentioning a software developer's "intended use" as a "very important factor" in determining its regulation); Cortez, *Digital Health*, *supra* note 12; Gerke, *supra* note 41, at 449, 468 (noting that intended use can free medical AI/ML systems from device regulation entirely).

72. Cortez, *Digital Health*, *supra* note 12. While developers occasionally create devices that require the formation of a new category, this is very rare.

73. *See generally* James M. Flaherty, Jr., *Defending Substantial Equivalence: An Argument for the Continuing Validity of the 510(k) Premarket Notification Process*, 63 FOOD & DRUG L.J. 901 (2008); *Premarket Notification 510(k)*, U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH, https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-notification-510k (last visited Aug. 22, 2024). While different technological characteristics are a criteria for determining classification, this is rarely invoked in the

category and assigned risk.[74] A Medical Image Analyzer that highlights boundaries between bodily structures using deep-learning algorithms is treated as having the same "risk" as one that functions by allowing its operator to apply basic adjustments to image brightness and contrast.

This category-internal insensitivity to technological characteristics and method of operation is comorbid with another problem: that many device categories are overbroad. Devices are often lumped together by intended use and indication even if they present objectively different risk profiles. As an illustrative example, consider the Therac-25 radiation therapy device. This device was designed to dose patients with therapeutic levels of electron beam or X-ray radiation. Incidentally, it was also capable of generating radiation doses hundreds of times in excess of therapeutic levels. A software glitch caused these machines to generate extreme doses on multiple occasions, resulting in several injuries and deaths.[75] Another device, the Uni-frame Head Immobilization Clamp (a simple mechanical clamp to restrict patient movement for radiation therapy procedures), is in the same device category as the Therac-25.[76] While a clamp failing to properly immobilize a patient during a radiation therapy procedure may have serious consequences, the consequences are ostensibly caused by radiation produced by the radiation therapy machine, not the mechanical aspects of the clamp. Despite this clear difference in risk profile, the two devices share a category and risk classification and are therefore subject to the same rigor of post-classification regulatory scrutiny.[77]

Further undermining the meaningfulness of risk classification is the inconsistent application of "special controls" in Class II risk product categories. Class II risk is defined by the inadequacy of Class I general controls to ensure device safety and efficacy, yet many, if not most, Class II product categories have no defined special controls at all. Even the product category that contained the lethal Therac-25[78] to this day has no defined special controls.

The introduction of AI "devices" has also further broadened the already overbroad scoping of device categories. The FDA has been allowing manufacturers to categorize AI devices into product categories that long predate

context of AI/ML devices. Almost all AI/ML devices are in product categories that were initially populated by non-AI/ML systems.

74. Discussed in more detail, *infra*, regarding the 510(k) substantial equivalence pathway.

75. Nancy G. Leveson & Clark S. Turner, *An Investigation of the Therac-25 Accidents*, 26 COMPUTER 18, 21–24 (1993).

76. Medtec's Uni-frame Head Immobilization System (K933227) is approved as a "Medical charged-particle radiation therapy system." Letter from Robert A. Ochs, Director, Center for Devices and Radiological Health, U.S. Food & Drug Admin., to Alena Newgren, Regulatory Specialist (June 29, 2018), https://www.accessdata.fda.gov/cdrh_docs/pdf18/K180021.pdf.

77. The CDRH has a separate set of regulatory requirements specifically for radiation-emitting products, but those requirements are parallel to the device risk classification and approval pathway.

78. There are separate FDA regulations dealing specifically with radiation-producing devices, but these do not appear to be robustly connected to device/risk classification. Other examples of non-radiation-related Class II devices without special controls are not difficult to find.

medical AI technology, without updating those category definitions with new special controls. This is a direct consequence of the "purpose" and "medical indication" tests for product categorization. As long as new AI devices fulfill a category's purpose and indication criteria, the technological risks presented by AI systems remain a non-factor.

At least one Class II category has bucked the trend, establishing comprehensive special controls that are at least relevant to evaluation of AI systems. However, the majority of AI-inclusive categories have not added any AI-specific special controls. Today, the five largest product categories of AI devices, comprising nearly 50 percent of all approved AI devices, have no AI- or ML-relevant special controls.

For a set of technologies as new as AI, we would expect risk classifications to skew higher than traditional devices, yet only 16 of the 1,247 approved AI devices (1.3%) were classified as Class III, with 98.7% classified as Class II.[79] If the risk classification system had any relation to common understandings of risk, technologies resistant to traditional methods of device evaluation (e.g., "black-box" AI/ML systems) should land higher on risk categorization, with a healthy proportion of Class III classifications.

The overwhelming skew of AI devices toward Class II could indicate that AI/ML developers preferentially choose to develop devices that fit defined Class II product categories to avoid the Class III premarket approval process. This choice is enabled by the broadness of categories and complete insensitivity of risk classification to technological risks. This insensitivity also flattens the spectrum of AI risk, creating absurd outcomes. For example, under the current classification system, SnoreSounds™ AI snoring analysis software (K150102) and the Brainomix AI rapid diagnostic software for assessing stroke progression within 6 hours of ischemic event (K243294) are both Class II devices, treated as presenting comparable levels of risk for the purposes of device approval. Detection of snoring can potentially be used as an indicator for sleep apnea,[80] but detection of an apnea-indicative snore rarely warrants the type of immediate critical medical attention that stroke detection requires. Even ignoring the AI characteristics of these devices, the fact that two devices are subject to the same

79. *AI Medical Devices*, *supra* note 3. Looking at approvals of all medical devices not exempt from premarket notification within the time period 2016-2024, 3.5% were categorized Class I, 73% as Class II, and 22% as Class III. Class II classifications clearly dominated approvals, but unlike for AI/ML device approvals, there was a nominal proportion of devices in Class I and a significant proportion in Class III.

80. Hui Jin, Li-Ang Lee, Lijuan Song, Yanmei Li, Jianxin Peng, Nanshan Zhong, Hsueh-Yu Li & Xiaowen Zhang, *Acoustic Analysis of Snoring in the Diagnosis of Obstructive Sleep Apnea Syndrome: A Call for More Rigorous Studies*, 11 J. CLIN. SLEEP MED. 765 (2015) (showing the potential utility of monitoring of snoring); B. Hofauer, B. Braumann, C. Heiser, M. Herzog, J.T. Maurer, S. Plößl, J.U. Sommer, A. Steffen, T. Verse & B.A. Stuck, *Diagnosis and Treatment of Isolated Snoring—Open Questions and Areas for Future Research*, 25 SLEEP BREATH. 1011 (2021) (discussing inconclusive results for snore monitoring in sleep apnea diagnosis).

level of regulatory scrutiny shows a fundamental disconnect between common understandings of risk and the FDA's treatment of risk.

Ideally, medical device risk classification would at minimum account for both the health-outcome criticality of the device's intended function(s), and the novelty and complexity of the mechanisms by which the device achieves its intended function(s). Novelty presents inherent uncertainty, and increased complexity warrants more rigorous documentation, testing, and evaluation, given that increased complexity increases potential failure modes.[81] ML models represent the current apex of novelty- and complexity-based risk. If an ML model is the critical determinant of the function of a medical device, that device should be treated as having enhanced risk by default.

B. 510(k): Rubber-Stamping Untested AI

The failure of the FDA's risk classification system enables what is perhaps the most critical problem with AI risk classification. By shunting nearly all AI devices into Class II risk, the FDA opens the door for device approval via the 510(k) pathway. The 510(k) pathway is designed to fast-track the approval of Class II devices with no requirements for testing, clinical or otherwise, as long as the manufacturer can demonstrate that their device is "substantially equivalent" to a previously marketed "predicate" device.[82] The bar for substantial equivalence has so far proven to be concerningly trivial for medical AI to pass, given that 96 percent of all approved AI-enabled devices were approved through 510(k).[83]

The 510(k) pathway was established by the 1976 MDA ostensibly to streamline approval of devices that were very similar to existing devices. This may have made sense for 1970s-era devices that were not nearly as complex as software (and AI) devices are now. Substantial equivalence can be measured against any relevant approved Class II device made by any manufacturer, not just previous models marketed by the applicant manufacturer. Furthermore, devices approved through 510(k) can serve as predicates for subsequent 510(k) applications. In practice, the 510(k) system has been incredibly permissive regarding what counts as "substantially equivalent." Most approved devices have extremely long predicate chains of consecutive 510(k) approvals, many stretching back to pre-1976 devices. To draw an analogy, 510(k) is the regulatory equivalent of a game of telephone, where the latest device in a predicate chain is usually unrecognizably different in mechanism from its original predicate approved decades ago. Given the lack of testing requirements in 510(k),

81. Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, 116 (2017) (mentioning the complexity of systems as reasons for more expert regulation). In a twist, this Article seeks to extend Tutt's thesis calling for an "FDA for algorithms" to call for an FDA for algorithms—for the FDA.

82. *See* Arianne Freeman, *Predicate Creep: The Danger of Multiple Predicate Devices*, 23 ANNALS HEALTH L. ADVANCE DIRECTIVE 127, 127–28 (2013).

83. 1195 out of 1247 devices were approved through 510(k).

problems in device safety and efficacy from small but compounding changes propagate through predicate chains in a process known as "predicate creep."[84] Devices can also have multiple immediate predicates, each covering a separate range of functionality. For example, the DePuy ASR XL hip replacement system had six immediate predicates, none of which covered the full set of functionality in combination.[85] Combining functions present new, unpredictable sources of risk, risks that are not considered by the 510(k) process.

Criticisms of 510(k) abound, generally centering on the facts that the 510(k) allows devices that have never been formally evaluated by the FDA for safety and effectiveness to serve as predicates,[86] and that the FDA allows new devices to be based on predicates demonstrated to be unsafe or ineffective.[87] The pathway received criticism as an overused loophole[88] within two years after its creation. Even the FDA has expressed awareness of the bad optics of allowing devices to be based on decades-old predicates. The FDA released a statement in 2018 expressing an intention to make public a list of all medical devices based upon a predicate greater than ten years old.[89] Of course, a list does not prevent old predicate use, nor does it address the fact, as some have pointed out, that devices with newer predicates are not necessarily any safer or more effective than devices with older predicates.[90]

The predicate chaining problem is particularly concerning for AI devices because AI devices are being predicated on non-AI devices. The figure below illustrates the predicate chain of an AI device called WRDensity (K202013) developed by Whiterabbit.ai and approved in 2020.[91] The product is marketed as using AI to interpret breast density imaging.[92] WRDensity is listed at the top of the figure, with predicates proceeding chronologically downward. The blue

84. *See* Freeman, *supra* note 82, 128.

85. George Horvath, *Empirically Assessing Medical Device Innovation*, 25 MINN. J.L. SCI. & TECH. 73, 109 (2023); Freeman, *supra* note 82, at 135.

86. *See e.g.*, Jonathan J. Darrow, Jerry Avorn & Aaron S. Kesselheim, *FDA Regulation and Approval of Medical Devices: 1976-2020*, 326 JAMA 420 (2021).

87. *See* Flaherty, *supra* note 73, at 919 (referring to statements by the Eleventh Circuit Court of Appeals in *Goodlin*, further noting a 20-hour average review turnaround for 510(k) applications).

88. Foote, *supra* note 65; Marilyn Uzdavines, *Dying for a Solution: The Regulation of Medical Devices Falls Short in the 21st Century Cures Act*, 18 NEV. L.J. 629 (2017).

89. U.S. FOOD AND DRUG ADMIN., OFFICE OF THE COMMISSIONER, STATEMENT FROM FDA COMMISSIONER SCOTT GOTTLIEB, M.D. AND JEFF SHUREN, M.D., DIRECTOR OF THE CENTER FOR DEVICES AND RADIOLOGICAL HEALTH, ON TRANSFORMATIVE NEW STEPS TO MODERNIZE FDA'S 510(K) PROGRAM TO ADVANCE THE REVIEW OF THE SAFETY AND EFFECTIVENESS OF MEDICAL DEVICES (2018), https://wayback.archive-it.org/7993/20190206212012/https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm626572.htm.

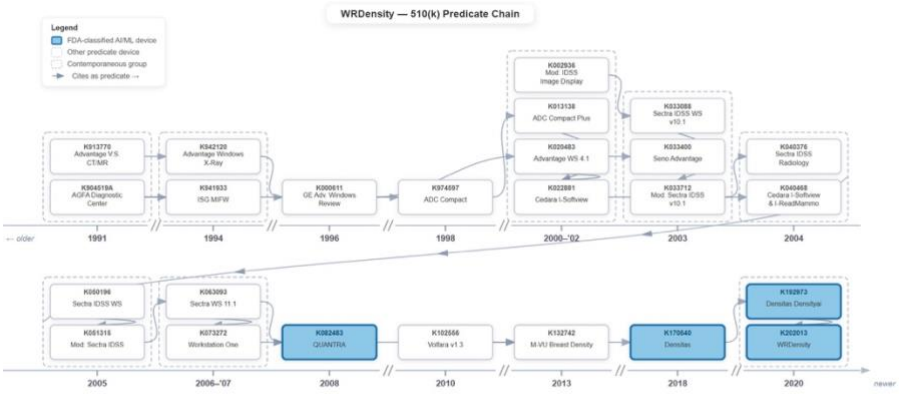
90. *See* Gerke, *supra* note 41, at 476–77.

91. 510(k) approval number K202013, product code QIH (Automated Radiological Image Processing Software). As a brief aside, this is one of the most common product categories for AI/ML devices, subject to special controls irrelevant to the operational mechanisms of the devices.

92. *WRDensity*, WHITERABBIT.AI, https://www.whiterabbit.ai/wrdensity (last visited July 29, 2024).

highlighted devices are considered AI-enabled by the FDA, while the others are not. This is just a partial predicate chain that doesn't show the full predicate history; the FDA did not retain predicate records for this device earlier than 1991.

Figure 2: Timeline of WRDensity's predicate chain. Devices highlighted in blue are classified by the FDA as AI/ML medical devices.



This pattern of predicating AI devices on non-AI devices is extremely common. Yet such predicate chaining should only be possible with an extremely permissive interpretation of the statutory definition of "substantial equivalence," from 21 U.S.C. § 360c:

> **(i)** Substantial equivalence
>
> **(1)(A)** For purposes of determinations of substantial equivalence under subsection (f) and section 360j(l) of this title, the term "substantially equivalent" or "substantial equivalence" means, with respect to a device . . . that the device has the same intended use as the predicate device and . . . the device—
>
> **(i)** has the same technological characteristics as the predicate . . . , or
>
> **(ii)(I)** has different technological characteristics . . . (II) does not raise different questions of safety and effectiveness than the predicate device.
>
> **(B)** For purposes of subparagraph (A), the term "different technological characteristics" means . . . that there is a significant change in the materials, design, energy source, or other features of the device from those of the predicate device.[93]

93. 21 U.S.C. § 360c(h)(i).

To summarize, a device can demonstrate substantial equivalence with a predicate in one of two ways: (1) having the same intended use and technological characteristics as the predicate device or (2) having different technological characteristics but the same intended use and demonstrably equivalent safety to the predicate device, without raising different questions of safety or effectiveness. Reviewers may use clinical or scientific data to support these findings, but only if they deem it necessary.[94]

The first method of demonstrating substantial equivalence—showing that a device has the same technological characteristics as its predicate—seems to be a reasonable standard by which to establish equivalence for physical and software medical devices, depending on the level of specificity. This type of evaluation does not work for AI systems. ML models, even those with identical architectures trained on the exact same data, have no guarantee of exhibiting similar performance. This is true for even seemingly minor adjustments. The most common method of "updating" medical ML models is by replacing one or more layers of a model with new, randomly initialized layers.[95] The idea is that retaining (and often "freezing") most of the existing layers of a model retains useful learned features from the previous model, while allowing the replaced, reinitialized layers to learn from new data. Training models this way requires far less computation, and therefore expense, than training models from scratch, but there is no guarantee that retraining a few layers of a well-established model will yield models that learn the same features. One study trained four identical models with a convolutional architecture from scratch with random initializations and discovered that while some learned features were reliably expressed across models, other learned features had no analogs in their twin models.[96] Another study demonstrated a similar result on another model architecture, an encoder transformer. Encoder transformers LLMs transform text into numerical representations of semantic meaning. One of the most widely used early encoder models, BERT (Bidirectional Encoder Representations from Transformers), is used frequently as a base model for retraining on specific textual domains (e.g., medical text).[97] The study tested this method of retraining on BERT multiple

94. *Id.* § 360c(h)(i)(D)(i) ("Whenever the Secretary requests information to demonstrate that devices with differing technological characteristics are substantially equivalent, the Secretary shall only request information that is necessary to making substantial equivalence determinations. In making such request, the Secretary shall consider the least burdensome means of demonstrating substantial equivalence and request information accordingly.").

95. Lotta M. Meijerink, Zoë S. Dunias, Artuur M. Leeuwenberg, Anne A.H. de Hond, David A. Jenkins, Glen P. Martin, Matthew Sperrin, Niels Peek, René Spijker, Lotty Hooft, Karel G.M. Moons, Maarten van Smeden & Ewoud Schuit, *Updating Methods for Artificial Intelligence-Based Clinical Prediction Models: A Scoping Review*, 178 J. CLIN. EPIDEMIOL. 111636 (2025).

96. Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson & John Hopcroft, *Convergent Learning: Do Different Neural Networks Learn the Same Representations?*, ARXIV 1, 5, 9 (Feb. 28, 2016), https://arxiv.org/abs/1511.07543.

97. The most pervasive use of encoder models is to find chunks of text with similar meaning in a large document library. Encodings with similar meanings are generally closer together. This applies to question- answering too; a question will generally have an encoding that is close to its answer. Jiajia

times with different random initializations of the final layer.[98] The study found that even reinitializing just the final layer of BERT yielded extremely high variances in model accuracy (spanning tens of percentage points). The "same technological characteristics" test for equivalence has no objective relevance to model function and performance and should not be a basis for a finding of substantial equivalence for medical AI systems.

As for the second method of establishing substantial equivalence—basing a device on a predicate with different technological characteristics—it is unclear by what standards the FDA determines whether a device (a) demonstrates at least equivalent safety to, or (b) raises "different questions of safety or effectiveness" from its predicate. The FDA's approval history of AI devices based on non-AI predicates seems to demonstrate that the FDA takes the optionality of reviewing supporting clinical or scientific evidence very seriously. Data bears this theory out: a 2024 study revealed that only 3% of approved AI devices included the disclosure of a clinical trial by the manufacturer.[99]

As written, the enabling statute for 510(k) is problematic for medical AI, given that similarity of technological characteristics is not relevant to the safety or effectiveness for AI systems. Some posit that the only way to fully address the question of safety for AI systems is through mandatory clinical testing and randomized controlled trials.[100] A 2022 review examining the performance of deep learning-based diagnostic radiology systems supports this contention.[101] The review compiled the accuracy of these AI radiology systems on data external to their training and validation sets. It found that 70 out of 86 systems exhibited decreased performance on external data. About half of the systems exhibited marginal accuracy decreases of 5% or less, while the others performed worse, with a few exhibiting over 20% reduction in accuracy. This review clearly shows that internal validation of model performance is almost always insufficient to predict real-world performance. Actual clinical testing and validation is critical for determining the safety and efficacy of medical AI.

Wang, Jimmy X. Huang, Xinhui Tu, Junmei Wang, Angela J. Huang, M.D. Tahmid Rahman Laskar & Amran Bhuiyan, *Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges,* 56 ACM COMPUTING SURVS. (2024).

98. Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova, *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, ARXIV (May 24, 2019), https://arxiv.org/abs/1810.04805.

99. Geeta Joshi, Aditi Jain, Shalini Reddy Araveeti, Sabina Adhikari, Harshit Garg & Mukund Bhandari, *FDA-Approved Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices: An Updated Landscape*, 13 ELECTRONICS 498, 509 (2024).

100. *See, e.g.* Thomas Y.T. Lam, Max F.K. Cheung, Yasmin L. Munro, Kong Meng Lim, Dennis Shung & Joseph J.Y. Sung, *Randomized Controlled Trials of Artificial Intelligence in Clinical Practice: Systematic Review*, 24 J MED. INTERNET RSCH. e37188, 2022.

101. Alice C. Yu, Bahram Mohajer & John Eng, *External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review*, 4 RADIOL.: A.I., 2022.

C. AI Off-Label: Systems Without Guardrails

Mandatory clinical testing of medical AI for its intended indications and purposes would go a long way in ensuring purpose-specific safety and efficacy, but medical AI faces another safety issue in that the medical device regulatory framework is not currently equipped to mitigate off-label uses of AI systems. Off-label use refers to the use of medical products in ways that are inconsistent with the purposes or indications for which the products were approved. For drugs, safe dosage thresholds established during mandatory clinical trials partly mitigate the danger of off-label use. Under device regulation, the complete lack of testing requirements means that analogous safe limits of AI device usage never have the opportunity to be evaluated.

What makes off-label use of AI systems uniquely risky is that, by dint of their ML nature, they often have inherent potential applicability that exceeds their approved purposes and indications. For example, the Brainomix 360 Triage Stroke[102] is an AI system approved for triage of medical images for indications of large vessel occlusion or intracranial hemorrhage (signs of stroke). The system's sole function is to send notifications to a physician, flagging suspected cases of stroke for priority human review. The fact that the model is used to flag images for signs of stroke, however, implies that the underlying ML model is trained for the task of diagnosis, even if it is not supposed to be independently used for diagnosis. Brainomix 360 Triage Stroke and devices like it generally do not provide highlighted or segmented images, nor any analysis to doctors beyond notification, but it is conceivable that "Triage" AI could end up being used off-label as a source of primary diagnosis rather than as simple triage. If, through use in practice, doctors find subjectively that a triage AI's notifications frequently correlate with positive diagnosis, they may eventually end up relying on it as de facto diagnostic criteria. Should such reliance become common, the off-label usage may enter general standards of care, without any empirical off-label study of reliability. Once a treatment enters the standard of care, its uses can persist for years before being found ineffective or dangerous.[103] Inappropriate reliance on non-diagnostic medical AI recommendation systems already show evidence of enhancing "correct" decision biases, while magnifying "incorrect" decision biases to a much greater degree in medical decision-

102. Letter from Jessica Lamb, Assistant Dir., Imaging Software Team, Ctr. for Devices & Radiological Health, to Brainomix Limited (Nov. 6, 2023), https://www.accessdata.fda.gov/cdrh_docs/pdf23/K232496.pdf (last visited July 19, 2024).

103. *See* George Horvath, *Off-Label Drug Risks: Toward a New FDA Regulatory Approach*, 29 ANNALS HEALTH L. & LIFE SCIS. 101, 102 (2020).

makers.[104] Some hypothesize that fear of liability may even encourage practitioners to rely more on AI outputs than their own judgment.[105]

An article by Meera Krishnamoorthy, Michael W. Sjoding, and Jenna Wiens[106] framed the problem of ensuring patient safety in the face of off-label use of AI as a problem of evaluative methods. The article defined two evaluative paradigms for AI systems: Empirical and Mechanistic. Empirical evaluations were evaluations in which AI models were tested in an off-label setting, with a sufficient volume of test cases to measure the AI model's effectiveness in the case of off-label use. Mechanistic evaluations were described as "qualitative evaluation of the difference between the off-label use and the approved use,"[107] and only possible when the decision-making processes of the AI system involved were well understood, or in other words, interpretable.

Figure 3: Krishnamoorthy's "spectrum of scientific support for off-label uses" of AI models[108]

	Weak evidence	Medium evidence	Strong evidence
Mechanistic evaluations	Weak understanding of the model's decision-making process and weak understanding of the similarity between approved and off-label use	Strong understanding of the model's decision-making process and weak understanding of the similarity between approved and off-label use, or vice versa	Strong understanding of the model's decision-making process and strong understanding of the similarity between approved and off-label use
Empirical evaluations	Retrospective evaluation at a site different from the off-label site	Prospective evaluation at a site different from the off-label site, or retrospective evaluation at a site similar to the off-label site	Prospective evaluation at a site similar to the off-label site

The article's discussion of evaluative methods is not particularly useful in providing solutions to the off-label AI problem, but the Empirical/Mechanistic

104. *See, e.g.*, Dane A. Morey, Michael F. Rayo & David D. Woods, *Empirically Derived Evaluation Requirements for Responsible Deployments of AI in Safety-Critical Settings*, 8 NPJ DIGIT. MED. (2025); Sarah Jabbour, David Fouhey, Stephanie Shepard, Thomas S. Valley, Ella A. Kazerooni, Nikola Banovic, Jenna Wiens & Michael W. Sjoding, *Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study*, 330 JAMA 2275 (2023).

105. See, e.g., Mahmut Alpertunga Kara, Clouds on the Horizon: Clinical Decision Support Systems, the Control Problem, and Physician-Patient Dialogue, 28 MED. HEALTH CARE & PHIL. 125 (2025). W. Nicholson Price II, Clinicians in the Loop of Medical AI, 74 EMORY L.J. 1265 (2025).

106. Meera Krishnamoorthy, Michael W. Sjoding & Jenna Wiens, *Off-Label Use of Artificial Intelligence Models in Healthcare*, 30 NAT. MED. 1525 (2024).

107. *Id.* at 1526.

108. Krishnamoorthy, *supra* note 106, at 1526.

dichotomy is a good framing for why medical device regulation does not work for off-label AI and what might work better.

Medical device risk assessment and 510(k) are both based on qualitative, mechanistic evaluations of devices. These types of evaluations work for devices that are interpretable in their form and function. Take, for example, the Peripherally Inserted Central Catheter (PICC). A PICC is, in basic terms, a tube inserted into a vein and directed through vasculature for long-term intravenous access. The FDA-approved parameters for PICC usage are very specific: PICCs must have a diameter between 3Fr and 6Fr; have a single, double, or triple lumen; must be inserted by modified Seldinger technique; and must be guided by ultrasound in the deep veins of the arm.[109] Approved method and indication aside, PICCs are used off-label with a huge variety of insertion techniques, insertion sites, and treatment indications. These off-label uses were evaluated by matching the physical properties of the PICC (e.g., diameter, number of lumens, pressure, flow rate) to the physical properties of the intended site of use. The PICC's structure and function clearly define the bounds of the PICC's potential safe uses.

Contrast the off-label uses of PICCs with those of a drug like Adderall. Adderall was approved for use in ADHD treatment, but empirical testing revealed loss of appetite as a side effect. One physician, Dr. Fuad Ziai, prescribed Adderall to approximately 800 children for the purpose of weight loss, a clinical target that had nothing to do with the intended mechanistic purpose of Adderall.[110] As with many other drugs, the clinically viable side effects were not predictable through mechanistic analysis, nor could reasonable safe dosages for other purposes be determined without empirical testing.

Off-label use of devices, drugs, and biologics is common, expected, and often essential in the practice of medicine. Such uses have value in both tailoring care and promoting innovation in healthcare delivery.[111] Many off-label uses have achieved acceptance as the standard of care for particular indications[112] and are sometimes the only way to use medical products to treat populations (e.g.,

109. Mauro Pittiruti, *The "Off-Label" Use of PICCs*, *in* PERIPHERALLY INSERTED CENTRAL VENOUS CATHETERS 127, 128 (Sergio Sandrucci & Baudolino Mussa eds., 2014), https://doi.org/10.1007/978-88-470-5665-7_12.

110. Madeline J. Cohen, *Off-Label: Combating the Dangerous Overprescription of Amphetamines to Children*, 82 GEO. WASH. L. REV. 174 (2013).

111. *See* David C. Radley, Stan N. Finkelstein & Randall S. Stafford, *Off-Label Prescribing Among Office-Based Physicians*, 166 ARCHIVES OF INTERNAL MED. 1021 (2006); Rachel E. Davis, *Regulating the Off-Label Use of Artificial Intelligence and Machine Learning-Enabled Medical Devices*, 26 VAND. J. ENT. & TECH. L. 771 (2023); Randall S. Stafford, *Regulating Off-Label Drug Use — Rethinking the Role of the FDA*, 358 NEJM 1427 (2008).

112. *See* Jennifer L. Herbst, *The Short-Sighted Value of Inefficiency: Why We Should Mind the Gap in the Reimbursement of Outpatient Prescription Drugs*, 2 CASE W. RES. J.L. TECH. & INTERNET 1 (2011).

pediatric populations), for whom clinical testing is not possible for ethical or practical reasons.[113]

Given the inherently experimental nature of off-label use, one might expect a higher frequency of negative outcomes than seen in practice. This dearth of negative outcomes is likely attributable to using the proper type of safety analysis for a given product type: mechanistic evaluative structures for interpretable medical devices and empirical evaluations for non-interpretable drugs. Shifting the evaluation method of non-interpretable AI systems from mechanistic to empirical will not guarantee safe off-label uses in isolation. Real-world clinical trials must be instituted for AI to study actual practitioner usage behaviors.

IV.
PATHS FORWARD FOR REGULATING MEDICAL AI

The current FDA approval framework for medical devices, while arguably adequate for traditional devices, is poorly suited to ensuring the safety and efficacy of medical AI systems. As discussed previously, this poor fit is attributable to (1) improper scoping of the term "medical device" to include intangible constructs like software and ML models, (2) effectively meaningless definitions of risk, (3) the 510(k) substantial equivalence loophole, and (4) the lack of effective safety evaluations for off-label uses of AI systems. Three potential paths of regulatory reform present themselves: (1) incremental, targeted reforms to medical device regulation that better accommodate the unique characteristics of AI systems, (2) reclassification of AI into the more suitable regulatory framework for drugs, and (3) creation of a new center within the FDA tasked with evaluating AI systems as a new class of medical product.

A. Incremental Reforms to Medical Device Regulation

Accommodating medical AI systems within the umbrella of medical device regulation has been the focus of the FDA and most AI policy scholars engaging with medical device regulation.[114] The framing of medical AI as "device" is mainstream.[115] AI is generally seen as a unique type of SaMD.[116] The FDA has

113. *See* David A. Simon, *Off-Label Innovation*, 56 GA. L. REV. 701, 740–42 (2021) (Simon suggests that incentivizing off-label data collection, such as by subsidizing healthcare providers who collect and compile such data in standardized ways, may help close the testing gap. Similar methods have worked in the past, like the Meaningful Use subsidy, a program that successfully incentivized broad adoption of Electronic Health Record systems.).

114. *See, e.g.*, Sara Gerke, *"Nutrition Facts Labels" for Artificial Intelligence/Machine Learning-Based Medical Devices—The Urgent Need for Labeling Standards*, 91 GEO. WASH. L. REV. 79 (2023); Charlotte A. Tschider, *Medical Device Artificial Intelligence: The New Tort Frontier*, 46 BYU L. REV. 1551 (2020); *supra* note 54.

115. *See, e.g.*, W. Nicholson Price II, *An Incidental Standard for Medical AI*, 8 J.L. & INNOVATION 1 (2025).

116. *See AI Medical Devices*, *supra* note 3 (Although approximately fifteen devices classified as AI-enabled devices predate the IMDRF definition of SaMD, AI-Enabled Medical Devices are a subcategory within SaMD according to the FDA webpage.).

attempted, and scholars have suggested, reforms to better ensure safety and efficacy of AI systems,[117] but such reforms tend to be limited by legacy structures like 510(k) that will be difficult to modify in a way that meaningfully addresses AI safety and efficacy without upending many of the existing systems of traditional device regulation.

For its part, the FDA has been hard at work altering device regulation to accommodate medical software and AI systems, but in a direction that seems uncertain to better address AI safety. The FDA is pushing for a Total Product Life Cycle (TPLC) paradigm of regulation.[118] This new TPLC approach ostensibly seeks to harmonize premarket review, postmarket surveillance, and compliance throughout a device's lifecycle by creating centralized databases for device information and new postmarket surveillance mechanisms, like the National Evaluation System for health Technology (NEST). TPLC also represents a shift away from premarket evaluation and towards postmarket monitoring and surveillance.[119] A significant catalyst for the TPLC approach was the need to accommodate the rapid proliferation of software medical devices.[120] Unlike physical devices, software development and maintenance require constant updates and patching. The idea is that a system of postmarket surveillance and monitoring would presumably accommodate software better than requiring new premarket applications for minor modifications to software. The FDA recently implemented Pre-determined Change Control Plans (PCCPs) for software and AI devices, allowing updates without new premarket filings.[121]

The FDA justified this shift away from premarket safety evaluation with a new, alternative philosophy of risk. Rather than seeing its role as protecting patients from untested medical devices before they enter the market, the FDA now considers "a patient's inability to access a potentially life-saving or life-sustaining treatment as a safety concern" and "may determine that its benefits outweigh the greater uncertainty about its risks if appropriate risk mitigations are taken (such as labeling)."[122]

117. *See* Gerke, *supra* note 41 (expressing some optimism for the 2017 pilot parallel performance-based pathway for AI, and arguing that AI Clinical Decision Support Software should be folded back into device regulation).

118. *See* TRANSPARENCY REPORT, *supra* note 5.

119. *See* Cortez, *Digital Health*, *supra* note 12, at 16.

120. U.S. FOOD & DRUG ADMIN., DIGITAL HEALTH INNOVATION ACTION PLAN (2017), https://www.fda.gov/media/106331/download (last visited July 18, 2025); Cortez, *Digital Health*, *supra* note 12.

121. *Recommendations*, *supra* note 35.

122. U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH, MEDICAL DEVICE SAFETY ACTION PLAN: PROTECTING PATIENTS, PROMOTING PUBLIC HEALTH 7 (2024), https://www.fda.gov/about-fda/cdrh-reports/medical-device-safety-action-plan-protecting-patients-promoting-public-health (The concept of labeling as an example of effective risk mitigation for an untested medical device seems of dubious value, if testing has not revealed what the warning should specify beyond a lack of testing.).

It is unclear whether the FDA's shift towards a TPLC framework will lead to a more robust safety monitoring environment or go the way of previous postmarket surveillance attempts. According to a 2024 report by the Government Accountability Office, the FDA is devoting resources to this new framework.[123] However, the FDA noted that there is a significant gap between its projected needs and the funding allocated by the government. Funding-related or not, the FDA has a history of poor follow-through on postmarket monitoring for medical devices. The FDA has run several such initiatives over the past few decades, including the Medical Device Reporting System, the Medical Product Safety Network, Sentinel, NEST, and the Manufacturer and User Facility Device Experience (MAUDE) database. All of these efforts have been plagued by incompleteness, inaccuracies, poor compliance, and inconvenient data structures.[124] Even the FDA itself states that given the poor data quality of medical device reports in MAUDE, the database is not intended to be used to evaluate rates of adverse events or establish cause.[125] While such history is not necessarily determinative of the future, the latest TPLC monitoring initiative seems to have a low probability of accomplishing its stated goals.

Given sufficient resources, however, a TPLC approach to medical AI (that supplements rather than replaces premarket testing) has potential merit. Software development and ML model building are uniquely susceptible to unexpected behavior and security vulnerabilities, especially given the intractability of reviewing every potential execution state of a piece of software and its component libraries.[126] After approval and during deployment, software and AI systems require constant testing, review, and adjustment to ensure safe operation. To address this constant need for patching and updating software under the TPLC framework, the Food and Drug Omnibus Reform Act of 2022 granted the FDA express authority to implement PCCPs for software and AI devices. The FDA, along with Canadian and U.K. agencies, subsequently released a guidance document outlining the process by which developers may update AI devices without a new premarket authorization, so long as detailed plans for forthcoming updates are provided in the device's initial marketing submission.[127]

123. *See generally* U.S. GOV'T ACCOUNTABILITY OFF., GAO-24-106699, MEDICAL DEVICES: FDA HAS BEGUN BUILDING AN ACTIVE POSTMARKET SURVEILLANCE SYSTEM (2024).

124. *See, e.g.*, Lisa Garnsey Ensign & K. Bretonnel Cohen, *A Primer to the Structure, Content and Linkage of the FDA's Manufacturer and User Facility Device Experience (MAUDE) Files*, 5 EGEMS at 12; William B. Feldman & Aaron S. Kesselheim, *Active Postmarketing Device Surveillance as a Legislative Priority*, 388 BMJ r484 (2025); Frederic S. Resnic & Sharon-Lise T. Normand, *Postmarketing Surveillance of Medical Devices—Filling in the Gaps*, 366 NEJM 875 (2012).

125. *About Manufacturer and User Facility Device Experience (MAUDE) Database*, U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH, (2024), https://www.fda.gov/medical-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities/about-manufacturer-and-user-facility-device-experience-maude-database (last accessed Apr. 9, 2025).

126. *See* Bryan H. Choi, *Crashworthy Code*, 94 WASH. L. REV. 39, 79–80 (2019).

127. CHANGE CONTROL PLANS, *supra* note 7; *Recommendations*, *supra* note 35.

This system of pre-determined software updates, while useful for bug-fixing for software, does not translate particularly well to updating ML models, as discussed in Section II.B. While industry generally views the application of PCCPs to modification of AI systems positively,[128] there is significant uncertainty as to what constitutes an update with the potential to affect the safety or efficacy of an AI model. Changing something as innocuous as the random noise pattern used to initialize a fine-tuning layer can cause drastic changes in model behavior and performance.[129] This is not a problem for software, given that the human in charge of updating generally knows specifically which features and functions are being updated and what the exact intended change in functionality should be.

Furthermore, while the FDA's guidance on premarket submissions encourages validation of AI-DSFs and provides examples of how to do so,[130] there are no standards or explicit requirements for validation practices. Thus far, the comprehensiveness and representativeness of validation studies that have resulted in AI/ML device approvals have varied widely, with over half of studies failing to report information as basic as sample size, and fewer than 2 percent including a published scientific study.[131]

The FDA's attempts to regulate AI devices have been at least responsive to the unique characteristics of AI systems, but there is little indication of whether these modifications (which take the form of additional reporting standards about AI model and training data characteristics)[132] will serve the FDA's mandate to ensure the safety and efficacy of AI (and other) devices. A significant obstacle to making AI devices safe under the medical devices framework is the limitations of the framework itself, most notably the 510(k) system of approval by substantial equivalence. The FDA piloted an alternative pathway for AI/ML device approval that is parallel to traditional 510(k) and a subset/modification of De Novo. This manufacturer pre-certification pilot program sought to partially substitute good company culture for traditional assessments of device safety.[133]

---

128. *See, e.g.*, Jacques Andre DuPreez & Olivia McDermott, *The Use of Predetermined Change Control Plans to Enable the Release of New Versions of Software as a Medical Device*, 22 EXPERT REV. MED. DEVICES 261 (2025).

129. *See id.*; Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi & Noah Smith, *Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping*, ARXIV (Feb. 15, 2020), https://arxiv.org/abs/2002.06305.

130. *See Lifecycle Management*, *supra* note 40.

131. Vijaytha Muralidharan, Boluwatife Adeleye Adewale, Caroline J. Huang, Mfon Thelma Nta, Peter Oluwaduyilemi Ademiju, Pirunthan Pathmarajah, Man Kien Hang, Oluwafolajimi Adesanya, Ridwanullah Olamide Abdullateef, Abdulhammed Opeyemi Babatunde, Abdulquddus Ajibade, Sonia Onyeka, Zhou Ran Cai, Roxana Daneshjou & Tobi Olatunji, *A Scoping Review of Reporting Gaps in FDA-Approved AI Medical Devices*, 7 NPJ DIGIT. MED. 1 (2024).

132. TRANSPARENCY REPORT, *supra* note 5.

133. *See* U.S. FOOD & DRUG ADMIN. CTR. FOR DEVICES & RADIOLOGICAL HEALTH, THE SOFTWARE PRECERTIFICATION (PRE-CERT) PILOT PROGRAM: TAILORED TOTAL PRODUCT LIFECYCLE APPROACHES AND KEY FINDINGS 2 (2022) [hereinafter *Software Precertification*], https://www.fda.gov/media/161815/download.

The FDA partnered with nine companies that it deemed excellent in several criteria, including quality, patient safety, clinical responsibility, cybersecurity responsibility, and proactivity. Software devices from these companies went through an alternative streamlined approval pathway. While this pilot program was met with some approval,[134] some senators criticized it as reaching beyond the FDA's statutory authority, a point that the FDA conceded at the end of the pilot program.[135] Without additional statutory authority, the FDA is limited in its ability to create alternative regulatory pathways that fall outside of traditional premarket approval, De Novo, and 510(k).

If medical AI systems must be kept under the ambit of medical devices, two interventions may be prudent: First, the language of FDCA section 513(i)(1)(A) could be altered to require that an AI device only claim a predicate if it confirms at minimum equivalent performance and reliability to the predicate device when tested on a current, recently compiled validation set not included in its training data. Such a test need not be a comprehensive clinical trial, but the testing protocol should demonstrate equivalent or better performance to a statistically significant degree.

Second, the Food and Drug Administration Modernization Act could be amended to allow the FDA to deny 510(k) submissions for uses other than the manufacturer's proposed intended use of the device. In many AI 510(k) submissions, the device is more capable than the manufacturer-defined "intended use" suggests. Allowing manufacturers of AI devices to market devices with off-label capability is dangerous because, unlike drugs, where the FDA can set toxicity thresholds for safe consumption, AI devices do not possess analogous criteria for untested off-label use. Additionally, if an AI device has functionality beyond the intended use, there is a very real danger that practitioners will use the additional functionality regardless of recommendations.

B. Category Switch: Reclassifying Medical AI as Drugs

The various efforts of the FDA and others to square medical AI with device regulation indicate a general recognition that medical AI systems are difficult to accommodate within the traditional structures of medical device regulation. Instead of trying to continue to force this fit, reclassifying AI systems as drugs may be a more practical solution.

Reclassification makes sense when considering the characteristics of articles regulated as drugs. Drugs are the original non-interpretable medical "black box."[136] The basic chemical properties of a drug are characterizable, but tracing out the full web of biochemical pathways a drug interacts with or disrupts is nearly impossible, and there is no guarantee of understanding even if it were

134. *See* Gerke, *supra* note 41; Tschider, *supra* note 114.

135. *Software Precertification*, *supra* note 133, at 13.

136. *See* Gerke, *supra* note 114, at 139 (mentioning that for decades, and possibly still, the mechanism of action of acetaminophen was/is unknown).

possible to do so.[137] Similarly, the structural architecture of an ML model is characterizable, but the magnitudes of learned parameters and nonlinear, entangled relationships between network elements created make interpretation of model features and mechanisms effectively impossible.

As with drugs, the only way to verify the safety and effectiveness of AI systems is to test them empirically. ML models are generally first validated against a portion of their training set left unused for the specific purpose of verifying that the ML model is learning representative features.[138] This unused validation set can also be used as a signal to stop training before the model starts overfitting (i.e. memorizing its exact dataset rather than generalizable features). [139] This validation by dataset segmentation, however, does not necessarily translate to real-world performance. If the training and validation set is pulled from a single source of data, overfitting on the unique characteristics of the single data source can still occur. The most reliable assessment of performance requires testing against data that is representative of the characteristics of the population the ML model will be used on, collected under real-world conditions.[140]

ML models require current, real-world data of their target populations, just as drugs require clinical trials that include members of their target populations. This is not to say that clinical trials for AI systems will necessitate the same format, phase structuring, and general expense as drug trials. Unlike drugs, many candidate models can be trained on and validated against the same datasets simultaneously. Organizations are developing Medical AI benchmarks to push for evaluative and comparative standards for medical systems, though as of yet there is no regulatory recognition of these.[141] ML Models can be discarded and retrained as many times as a developer's compute budget can accommodate. Most importantly, for most medical AI applications, tests need not require the

137. *See* Martin Lukačišin & Tobias Bollenbach, *Emergent Gene Expression Responses to Drug Combinations Predict Higher-Order Drug Interactions*, 9 CELL SYS. 423, 423 (2019); Elif Tekin, Casey Beppler, Cynthia White, Zhiyuan Mao, Van M. Savage & Pamela J. Yeh, *Enhanced Identification of Synergistic and Antagonistic Emergent Interactions Among Three or More Drugs*, 13 J. R. SOC. INTERFACE (2016).

138. Gael Varoquaux & Olivier Colliot, *Evaluating Machine Learning Models and Their Diagnostic Value*, *in* 197 MACHINE LEARNING FOR BRAIN DISORDERS 614 (Olivier Colliot ed., 2023), https://www.ncbi.nlm.nih.gov/books/NBK597473/.

139. Xue Ying, *An Overview of Overfitting and Its Solutions*, 1168 J. PHYS.: CONF. SER. 022022 (2019).

140. Krishnamoorthy et al., *supra* note 106.

141. *See, e.g.*, Alexandros Karargyris, Renato Umeton, Micah J. Sheller, Alejandro Aristizaba, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, Prakash Narayana Moorthy, Alexander Chowdhury, Junyi Guo, Sahil Nalawade, Jacob Rosenthal, David Kanter, Maria Xenochristou, Daniel J. Beutel, Verena Chung, Timothy Bergquist et al., *Federated Benchmarking of Medical Artificial Intelligence with MedPerf*, 5 NATURE MACH. INTEL. 799 (2023); *Stanford Develops Real-World Benchmarks for Healthcare AI Agents*, STANFORD UNIV. HUMAN-CENTERED A.I. (Sep. 15, 2025), https://hai.stanford.edu/news/stanford-develops-real-world-benchmarks-for-healthcare-ai-agents (last visited Apr. 14, 2026).

constant participation of test subjects; if relatively recent real-world data on a representative population and condition exists, data need not be recollected for every ML system directed at that population and condition. Sharing of curated datasets for AI training has become commonplace on platforms like Hugging Face.[142] The same could potentially be done for medical data, under the supervision of the FDA or another standards body. Of course, clinical trials of ML models will require at least some component of real-world clinical application by practitioners for final safety and efficacy evaluation.

Of course, medical AI systems are not a perfect fit for drug regulation for reasons of institutional expertise, efficiency, and long-term fit. There may be a high-level conceptual alignment between AI systems and drugs, but the practical differences between drugs and AI will add complications. First, CDER would have to invest in institutional expertise in AI, an investment that would only be useful in a siloed set of AI-related concerns. In contrast, the current regulator of AI-enabled devices, CDRH, regularly deals with all types of technology; skills and knowledge would likely be more transferable across the center. Additionally, dealing with AI's practical relationship with software could be problematic. Even if software algorithms and nonlinear statistical models are conceptually different "kinds," software enables the training and inference of AI. Every new medical product involving an AI system would have to be first considered a combination product, unless CDRH relinquished software to CDER, which makes little sense, given that software does not share any regulable properties with drugs. CDER could potentially be assigned authority over only those software systems that interact with AI, but this would result in regulatory duplication of concerns and could potentially shift more software from CDRH to CDER over time. Drug regulation of AI will also be muddied by software's update problem. While evaluating each new version of an ML model with a new Investigational New Drug Application is arguably the proper way to ensure AI safety and efficacy, software updates to training and inferencing algorithms may necessitate awkward regulatory carveouts. Finally, regarding long-term regulatory fit, there is no guarantee that AI technologies will evolve in a direction that will share essential characteristics with drugs. Current types of AI systems will probably remain, but the likelihood that future AI technologies will invoke entirely new sets of regulatory concerns within the next couple of decades is very high.

C. A New Center for Medical AI Evaluation and Research

While both CDRH and CDER could work in the short term, neither is likely to be a good long-term regulatory home for medical AI technologies. The main problem with homing medical AI within either center is that both centers are

142. Hugging Face is an example of the potential for collaborative sharing in the open-source tradition, not as a place to get reliable or privileged medical data. *The AI Community Building the Future.*, HUGGIING FACE (last visited Apr. 14, 2026), https://huggingface.co/.

primarily in the business of regulating medical *products*. AI technologies are evolving rapidly, with existing medical AI systems already blurring the lines between products and practitioners.

For example, RayStation,[143] an AI-enabled software radiation therapy device approved in March 2025, processes radiological images to generate 3D models of a patient's internal organ systems and uses that model to autonomously plan and robotically execute proton beam destruction of identified tumor tissue.

Is it proper to classify an AI system that reads MRIs, generates treatment plans, calculates therapeutic dosages, and then executes that plan as a "device"? Is a physician's action of clicking an "execute treatment" button a transfer of responsibility from the AI system or manufacturer to the physician or not? Does the FDA have regulatory authority over the treatment decisions made by AI, or does that fall outside of the FDA's purview as practice of medicine? Sometime in the not-too-distant future, the FDA will have to address these and similar questions. A new center dedicated only to medical AI technologies would be best positioned to evaluate and respond to rapid changes in AI technology and statutory directives.

In the wake of significant FDA workforce cuts in 2025,[144] opening and staffing a new center is likely not high on the FDA's list of priorities. The Trump administration, however, recently announced new priorities under America's AI Action Plan, specifically tasking the FDA and SEC with enabling the creation of "regulatory sandboxes" where "researchers, startups, and established enterprises can rapidly deploy and test AI tools while committing to open sharing of data and results."[145] The creation of a new center devoted specifically to testing and evaluating AI technologies may be surprisingly politically viable, given that the Action Plan also includes initiative goals like "Build an AI Evaluations Ecosystem," "Build World-Class Scientific Datasets," and "Advance the Science of AI."[146]

A new center—let's call it the Center for Medical AI Evaluation and Research (CMAIER)—could present the opportunity not only to establish meaningful standards for safety and efficacy of AI systems, but also to promote innovation as consonant with—rather than in tension with—the pursuit of safety and efficacy. The center, in line with AI Action Plan priorities, could curate medical training datasets and create benchmarking resources in collaboration

143. Letter from Lora D. Weidner, Assistant Dir., Radiation Therapy Team, Ctr. for Devices and Radiological Health, to RaySearch Laboratories AB (Apr. 4, 2025), https://www.accessdata.fda.gov/cdrh_docs/pdf24/K240398.pdf (last visited Aug. 4, 2025).

144. Ahmed Aboulenein & Maggie Fick, *Trump Layoffs Begin to Erode FDA Drug Review System*, REUTERS, Apr. 4, 2025, https://www.reuters.com/business/healthcare-pharmaceuticals/trump-layoffs-begin-erode-fda-drug-review-system-2025-04-04/.

145. EXEC. OFF. OF THE PRESIDENT, AMERICA'S AI ACTION PLAN 5 (2025), https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.

146. *Id.*

with other government agencies hosting open datasets, including NSF, NIH, CDC, and NIST. Perhaps there would also be an opportunity not only to host datasets, but also to establish libraries with CMAIER-validated open medical models created by universities and technology companies, such as Google's MedGemma model.[147]

The most critical reason for the creation of a center like CMAIER, however, is the category-defying potential of AI technologies. Medical AI has already been separated into regulated medical devices and unregulated clinical decision support software, a distinction that is unclear and could either sharpen or further blur as AI technology develops. AI technologies may also diverge in nature (particularly with the future possibility of adaptive AI, and possibly even licensed practitioner AIs), creating categories that require more granular differentiation than CDRH has been able to provide. Medical AI has the potential to grow into a category containing as broad and diverse a set of concerns as those already regulated by CDER or CDRH today.

V.
CONCLUSION

The FDA's primary mandate with respect to medical devices is to protect the public by ensuring that every device that passes FDA review is safe and effective. While the regulatory systems created to carry out this mandate may have worked for the pre-digital devices of the late 1970s and their analog successors, the FDA's inability to recognize and adapt to critically divergent aspects of new technologies, like software and ML, has resulted in the failure of device regulation to stop an accelerating flood of effectively untested, unvalidated AI technologies from entering medical practice. Over 1,200 AI-enabled devices have slipped through the 510(k) regulatory sieve, claiming equivalence with technologies that share category and purpose with their successors, but little else.

Instead of shoring up regulatory walls around AI in response, the FDA has enabled rubber-stamping AI device approvals, leaving the rigor and validity of manufacturers' internal testing practices underexamined. The FDA has also been opening the AI approval floodgates wider by pursuing a Total Product Lifecycle framework, weakening premarket review not just for AI systems but for all medical devices.

In many respects, the FDA has changed very little in the past forty years. The FDA of 1983 approved the Therac-25 radiation therapy device via 510(k) and did not meaningfully change software regulation even after the Therac-25's software glitch lethally irradiated several patients. Fast forward to the FDA of

147. GOOGLE RESEARCH & GOOGLE DEEPMIND, MEDGEMMA TECHNICAL REPORT (July 15, 2025), https://arxiv.org/abs/2507.05201.

2025 that has just approved RayStation, a high-tech AI cousin of Therac-25, via 510(k)—we can only hope that the similarity ends there.

Meaningful reform of medical AI regulation will require more than incremental adjustments to medical device regulation; real change will require the FDA to finally recognize the essential differences between devices, software, and AI. Forty years of treating revolutionary, category-defying technology as Brannigan's "new kind of bedpan" precipitated the current, fundamentally broken system of medical AI regulation. FDA leadership will recognize the need for action before a preventable AI-driven tragedy occurs. The current presidential administration's zeal for AI development could be the perfect opportunity for the FDA to reboot AI regulation and reaffirm its commitment to the safety and efficacy of medical AI technologies.